NATURAL LANGUAGE PROCESSING AND DEEP LEARNING APPROACHES FOR SUSTAINABILITY AND INFRASTRUCTURE POLICY ANALYSES

A Dissertation
Presented to
The Academic Faculty

by

Sooji Ha

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Computational Science and Engineering

School of Civil and Environmental Engineering
College of Engineering

Georgia Institute of Technology
August 2021

NATURAL LANGUAGE PROCESSING AND DEEP LEARNING APPROACHES FOR SUSTAINABILITY AND INFRASTRUCTURE POLICY ANALYSES

Approved by:

Dr. Emily Grubert, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Duen Horng Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Omar Isaac Asensio, Co-advisor
School of Public Policy
Georgia Institute of Technology

Dr. Chao Zhang
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Iris Tien
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: July 13, 2021

To my mom, whom I dearly miss, for inspiring me to dream big and instilling in me the resilience to pursue those dreams.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisors, Dr. Emily Grubert and Dr. Omar I. Asensio, for their mentorship and support during my journey as a PhD student. I can dare say that I had the most amazing experiences with my advisors who showed me not only their academic excellence but also their kindness, passion, and integrity, and set an exemple for me. Their guidance made my PhD journey such an enjoyable and memorable experience. I also wish to thank my committee members, Dr. Iris Tien, Dr. Chao Zhang, and Dr. Polo Chau for their professional guidance and insightful comments on my research.

I am also grateful to my lab members from the Grubert Group and the Data Science and Policy Lab. I especially would like to thank all the co-authors of my publications and those who helped me in the data collection processes. I also would like to extend my gratitude to my friends at Georgia Tech. My time at Georgia Tech will always remain such a happy memory thanks to all the boardgame nights and trips that we had together.

Most of all, I am extremely appreciative to my family - my dad who supported me throughout my academic journey, and my sister Sua and brother-in-law, Tom, who went out of their way to help me with my PhD journey from the beginning to the end. My sister's family provided me with a home and unconditional love that I could always turn to. Lastly, I would like to thank my husband, Sam, for accompanying me all the time and helping me go through all my lows and highs. I would not have completed my PhD degree without his heartwarming kindness and care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Public comments are essential components in the regulatory process, with ability to affect government's rulemaking process. Recently, unsolicited public comments from social media on mobile platforms, have opened new possibilities of engaging the public in government work. Recent computational advances in natural language processing (NLP) and deep learning have shown capabilities for utilizing these types of comments data for policy decision making process. However, when it comes to sustainable infrastructure policy domain, large amounts of publicly available comments data exist but they are yet to fully take advantage of the computational advances. Traditional methods on engaging public opinions for policy implementations include survey and interview processes. However, these survey-based approaches have major limitations as they are often slow and costly to collect. Computationally, unsupervised methods like topic modeling for exploratory analysis of unlabeled texts has been often used. However, such approaches have limitations on creating targeted, theoretically meaningful clusters, and they are not suitable for hypothesis testing, spatial analysis or benchmarking with other corpus.

This dissertation uses NLP with deep learning to overcome these challenges in engaging publicly available comments data with policy making processes. This dissertation expands the literature by demonstrating a framework of designing machine learning lifecycles tailored for sustainable policy analyses. First, I collect and analyzes government solicited public comments towards an energy policy. Next, I assess electric vehicle infrastructure system across the United States using consumer-generated social data using transformer models. Finally, I use attention flow quantification method to interpret the

transformer models and examine the model behaviors on predicting the topics of the user generated reviews.

Collectively, this dissertation demonstrates the potential of using public-generated comments and deep learning for research on the sustainable policy analysis and for discovering hard-to-reveal patterns in unstructured large-scale data that can provide useful insights to public policy advisory. The theoretical and methodological contributions of this dissertation help policy makers and industry experts understand the interactions between the public and infrastructure, and make targeted interventions that are effective and equitable.

CHAPTER 1. INTRODUCTION

Public comments are essential components in the regulatory process, with ability to affect government's rulemaking process (Costa et al., 2018, Ervin et al., 2019, Boustead and Stanley 2015). Solicited public comments provide feedback on proposed rulemakings, regulations, and updates, helping the government agencies consider diverse points of view and improve the quality of their policymaking. Recently, unsolicited public comments, such as from social media on mobile platforms, have opened new possibilities of engaging the public in government work, fostering interactions between policy makers and citizens (Lampe et al., 2011; OECD 2014; Lee and Kwak 2012).

Recent computational advances in natural language processing (NLP) and deep learning have shown capabilities for utilizing these types of comments data for policy decision making process. However, when it comes to sustainable infrastructure policy domain, large amounts of publicly available comments data exist but they are yet to fully take advantage of the computational advances, or even remain largely dormant.

Traditional methods on engaging public opinions for policy implementations include survey and interview processes. However, these survey-based approaches have major limitations as they are often slow and costly to collect, are limited to regional sampling, and are often subject to self-report or recency bias (Grubert, 2017; Walsh et al., 2020; Ha et al., 2021). Computational approaches for evaluating of high-volume public opinion corpora and social media data in this domain, has often used unsupervised methods like topic modeling for exploratory analysis of unlabeled texts (Blei, 2012; Hemmatian et al., 2019; Ma et al., 2016;

Wang et al., 2019). However, such approaches have limitations on creating targeted, theoretically meaningful clusters, and they are not suitable for hypothesis testing, spatial analysis or benchmarking with other corpus (Asensio et al., 2020a; Ha et al., 2021).

This dissertation uses NLP with deep learning to overcome these challenges in engaging publicly available comments data with policy making processes for greenhouse gas emission reduction. This dissertation expands the literature by demonstrating a framework of designing machine learning lifecycles tailored for sustainable policy analysis, that encompasses project objectives definition, data collection, conducting human experiments for training data curation, model training, and interpreting the model behaviors.

In Chapter 2, government solicited public comments towards an energy policy are collected and analyzed, using convolutional neural network with word embeddings and data augmentation. It finds that majority of the commenters supported the policy while it was ultimately repealed, with bifurcated arguments on justice topic, discussing just transition for fossil fuel industry workers and environmental justice.

In Chapter 3, the dissertation assesses electric vehicle infrastructure system across the United States using consumer-generated social data using transformer models. It finds that many micropolitan statistical areas could be underserved regarding charging station availability.

In Chapter 4, the dissertation uses attention flow quantification method to interpret the transformer models and examine the model behaviors on predicting the topics of the user generated reviews. The results show that the models effectively capture domain specific terms that are not easily recognized by general crowds.

Collectively, this dissertation demonstrates the potential of using public-generated comments and deep learning for research on the sustainable policy analysis and for discovering hard-to-reveal patterns in unstructured large-scale data that can provide useful insights to public policy advisory. The theoretical and methodological contributions of this dissertation help policy makers and industry experts understand the interactions between the public and infrastructure, and make targeted interventions that are effective and equitable.

CHAPTER 2. UNDERSTANDING THE PUBLIC ATTITUDES TOWARDS THE CLEAN POWER PLAN USING NATURAL LANGUAGE PROCESSING AND DEEP LEARNING[1]

2.1 Introduction

The Clean Power Plan (CPP), a measure proposed under the authority of the Clean Air Act, Section 111(d) to reduce U.S. greenhouse gases (climate change pollution) from electricity generating power plants, was proposed by the Obama administration in 2014 (final rule: 2015), with repeal proposed by the Trump administration in 2017 (US EPA, 2020). The CPP was formally repealed in 2019 without ever taking effect (US EPA, 2020). Public comment was solicited both when the measure was proposed, and when its repeal was proposed, with a total of more than 6 million comments submitted in total (General Services Administration, 2021). Most of the comments submitted are online and mail submissions, but these included about 2,000 spoken comments transcribed during public hearings and listening sessions in eight geographically, socially, economically, and politically diverse locations in the U.S.: Atlanta, GA (pre-implementation); Charleston, WV (pre-repeal); Denver, CO (pre-implementation); Gillette, WY (pre-repeal); Kansas City, MO (pre-repeal); Pittsburgh, PA (pre-implementation); San Francisco, CA (pre-repeal); and Washington, DC (pre-implementation) (US EPA, 2020). Notably, these locations include cities in places ranging from highly climate-regulated [e.g., San Francisco

[1] This chapter was submitted for review to Environmental Science and Technology journal by the American Chemical Society with Emily Grubert as the co-author.

(California Renewables Portfolio Standard Program: Emissions of Greenhouse Gases., 2018)] to highly coal dependent [e.g., Charleston, WV; Gillette, WY (Bell and York, 2010; Godby et al., 2015)], providing an opportunity to evaluate public attitudes toward an electricity-focused climate policy expected to have particularly significant impacts on coal country (Godby and Coupal, 2016).

Gathering public opinion data about major policy implementations like the proposed CPP and its repeal can be challenging, especially given tight timelines for acquiring funding, planning data collection efforts, and collecting such data during ongoing policymaking processes in the context of declining response rates (Czajka and Beyler, 2016; Stern et al., 2014). Spatial and temporal variability in public opinion poses interpretability challenges for snapshot data collected by means of surveys or brief interviews, particularly in contexts where conditions change rapidly (e.g., renewable energy prices; policy proposals). Human subjects data collection can be expensive, time consuming, and a source of research fatigue for communities of special interest for specific research topics (Grubert, 2017; Walsh et al., 2020). As such, public comments on policy implementations are extremely valuable sources of information about how the most engaged citizens react to specific policies, particularly when such comments are made in person and thus less likely to be form letters, bots, or other less individualized texts (De'Arman, 2020).

Public comments on the CPP are an especially useful example of public comment data as public opinion data for two major reasons. First, comments are available for both the pre-implementation and pre-repeal periods, with only a few years' separation, and entirely within the window of digital record keeping. Second, in-person listening sessions and hearings were geographically diverse. In the case of the pre-repeal sessions, these

sessions included small communities likely to be disproportionately affected by the CPP in an unusually targeted way, potentially biasing comments from a federal policymaking perspective but also allowing access to comments from areas where opinion data can be difficult to collect (Grubert, 2019). As such, public comment data associated with the CPP are a rich source of information on American opinion about both the CPP and climate policy more generally.

One major challenge with evaluating public comment data is the volume. Although data are publicly available and often transcribed, federal actions can receive thousands of comments (Shapiro, 2008). These comments are structured in the sense that they address the same general topic area, but they are generally not as structured as designed human subjects data like that from a survey or interview process. As such, manually analyzing public comments poses major readability and analytical challenges, including time and cost constraints (Haynes-Maslow et al., 2018). Although agencies evaluate and respond to comments, formal responses are highly bureaucratically inflected and do not typically include a goal of capturing public opinion (Costa et al., 2019; De'Arman, 2020). Computational approaches potentially offer a major opportunity to classify and evaluate large volumes of data (Grubert and Siders, 2016; Scott et al., 2020), with long-standing efforts to apply such tools to public comments (Xu and Bengston, 1997; Yang and Callan, 2009).

Prior work on computational approaches to describing and evaluating the content of high-volume public opinion corpora has often used unsupervised methods like topic modeling for exploratory analysis of unlabeled texts (Blei, 2012; Hemmatian et al., 2019; Ma et al., 2016; Wang et al., 2019). Although such computational methods have often been

applied to social media and review comments, these methods are less commonly used with the smaller but more targeted corpora associated with formal public comments, meeting transcripts, and similar opinion records. This is partly because unsupervised machine-learning methods, such as topic modeling, aim to provide a generalized summary of data rather than optimized answers to specific questions (Hemmatian et al., 2019), posing limitations on creating targeted, theoretically meaningful clusters (Ha et al., 2021). Topic classification on these types of documents using supervised methods like neural networks answer more targeted questions, but they require large amount applying conventional, manual coding techniques to a manageably-sized subset of labeled data as inputs (LeCun, 2015). We suggest that by applying conventional, manual coding techniques to a manageably-sized subset of an overall corpus, manual coding can be used to generate a training set for supervised machine learning that enables meaningful evaluation of much larger datasets.

This research employs a natural language processing (NLP) approach with a supervised deep learning algorithm called a convolutional neural network (CNN) as a multi-class and multi-label classifier to evaluate public opinion on the proposed CPP and its repeal, with the overall goal of understanding spatiotemporal variation in public attitudes about this specific climate policy. The remainder of this article describes our manual and computational methods, presents a validation of the computational modeling work, then presents and interprets model results for public opinion on the CPP during both the pre-implementation and pre-repeal periods, by city. Specifically, we evaluate measures of policy support, describe content themes present in the public comments, and assess

correlation between specific themes. We conclude with a description of limitations and future work.

2.2 Methods

To conduct descriptive analysis of public attitudes toward the CPP both pre-implementation and pre-repeal by location of public hearings and listening sessions, based on publicly available public comments, we combine manual coding and computational classification methods. This section describes our data sources and collection protocol, then the manual and computational analytical approaches.

2.2.1 Data Sources and Collection

CPP public comment data from in-person hearings and listening sessions were collected from regulations.gov under docket numbers EPA-HQ-OAR-2013-0602 and EPA-HQ-OAR-2017-0355 for pre-implementation and pre-repeal sessions, respectively. Collected comments were organized by city for use with our geographic frame: Atlanta (GA); Washington, DC; Pittsburgh (PA); and Denver (CO) for pre-implementation, and Kansas City (MO); San Francisco (CA); Gillette (WY); and Charleston (WV) for pre-repeal several years later. In these hearings, public comments were collected in three forms: spoken testimony, written testimony submitted at the hearing, and public hearing cards. This study relies on transcribed spoken testimony only, as these comments are available as digital text through the federal repository but public hearing card comments are not. After downloading documents from the repository, transcribed spoken testimony (available as a

single document for each session) was parsed by speaker using the following steps: First, the document was parsed by column signs (:) that followed the speakers' names, which occurred whenever the speakers changed. Second, **e**xcluding the starting remarks of the hearing chair, text chunks of 150 words or longer were selected. Through these two steps, the document successfully filtered out trivial conversations that were not parts of the comments. After parsing, the texts were cleaned up by removing numbers, excessive blank spaces, and page footers. The texts were then converted to lower cases. After cleaning the data, the average number of words was 619 per comment, with 4% of the comments exceeding 1,000 words that were from speakers given excessive time for speaking. For computational performance for model implementation, these comments were truncated at 1,000 words (See et al.,2017). Then, the average number of words was 584, with a minimum of 149 words. After pre-processing, our dataset comprises a total of 1,900 public comments, with distribution by location shown in Table 2.1.

Table 2.1 Distribution of comment data per hearing or listening session location

	Pre-Implementation				Pre-repeal				
	DC	Denver	Atlanta	Pittsburgh	Charleston	San Francisco	Gillette	Kansas City	Total
No. of data	399	391	363	179	171	163	162	72	1,900
Avg. word count (after truncation)	603	607	578	655	674	454	471	543	584

2.2.2 Manual coding: Typology development and training data curation

Two coders (the authors) manually reviewed a subset of the public comments described above, then iteratively developed a codebook and coded a stratified random

sample of the comments, with 25 comments from each of the eight public hearing or listening session locations. As these coded comments were intended both to support an overall typology for the data and to serve as training data for the computational approach described below, the decision to code a total sample of 200 comments was based on a target training dataset size of approximately 10% of the total dataset. The sample was divided equally across locations rather than proportional to total comments from each location due to this study's geographic frame and interest in identifying thematic diversity. Although the authors did not review every comment in the dataset, the total coded sample of 200 comments was sufficient for conceptual saturation at the coarse thematic level used here. Our assessment that 25 comments (effectively, focused mini-interviews) per location was sufficient for saturation is consistent with prior findings from that saturation is often reached well below 20 interviews (Guest et al., 2020). In addition to engaging the public comments themselves, one of us (Grubert) attended the Gillette and San Francisco sessions in person, with observations from those experiences reinforcing our assessment of conceptual saturation from our coded sample.

The codebook was developed with two separate categorical goals: 1) to identify explicit support for or opposition to implementing the CPP, and 2) to identify themes in commenters' statements providing explanatory support for this support or opposition. For the first goal, support or opposition for the CPP implementation was coded as multi-class classification, where each comment is assigned to one and only one label. To account for comments that are unrelated to CPP or that show no explicit support or oppose towards the plan, neutral class was included. For the second goal, multiple topics could be assigned to

each comment as commenters mention various themes for their statements. Therefore, it is coded as multi-label classification.

The authors independently reviewed comments and proposed themes, then collaboratively defined initial codes that were iteratively refined over three coding passes. Given the goal of using coded data as a labeled training set for NLP, after each pass, the authors assessed inter-rater reliability and then collaboratively reviewed all disagreements until perfect agreement was achieved. Pre-reconciliation, the average Cohen's kappa score ranged from 0.49 to 0.89 across 14 codes.

Final codes included a support/oppose measure and four high-level thematic topics with thirteen subtopics. Support or opposition was coded numerically, with 1 representing support for implementing the CPP (or opposition to repeal); 0 representing a neutral comment (e.g., off-topic comments that did not address the CPP); and -1 representing opposition to implementing the CPP (or support for repeal). High-level thematic topics of *Environmental Impacts*, *Economy*, *Resources*, and *Ethics* were further divided into thirteen subtopics: *Environmental Impacts* has *Climate, Health, Pollution,* and *Extreme Events* as subtopics; *Economy* has *Jobs, Costs,* and *Future Economy* as subtopics; *Resources* has *Coal, Natural Gas,* and *Clean Energy*; and *Ethics* has *Future Generations, Justice,* and *Stewardship*. The codebook, which includes a brief description of each code, guidance on how to implement the codes, explicit guidance on what is and is not included, and examples of correct and incorrect application where relevant, is available in Appendix A.

2.2.3 Computational classification: Convolutional Neural Network (CNN) modeling

For computational analysis of the full comment sample (1,900 comments), we employed a supervised machine learning approach because of our interest in descriptive rather than fully exploratory analysis. Specifically, we adapted a 1-layer convolutional neural network (CNN) model for both multi-class and multi-label tasks, using word2vec word embeddings, following the protocols suggested by Kim (2014) and Asensio et al. (2020a). Unlike unsupervised approaches like topic modeling, CNN models require training with labeled data: in this application, because the input data were initially unlabeled, we curated our own training data via the manual coding described above.

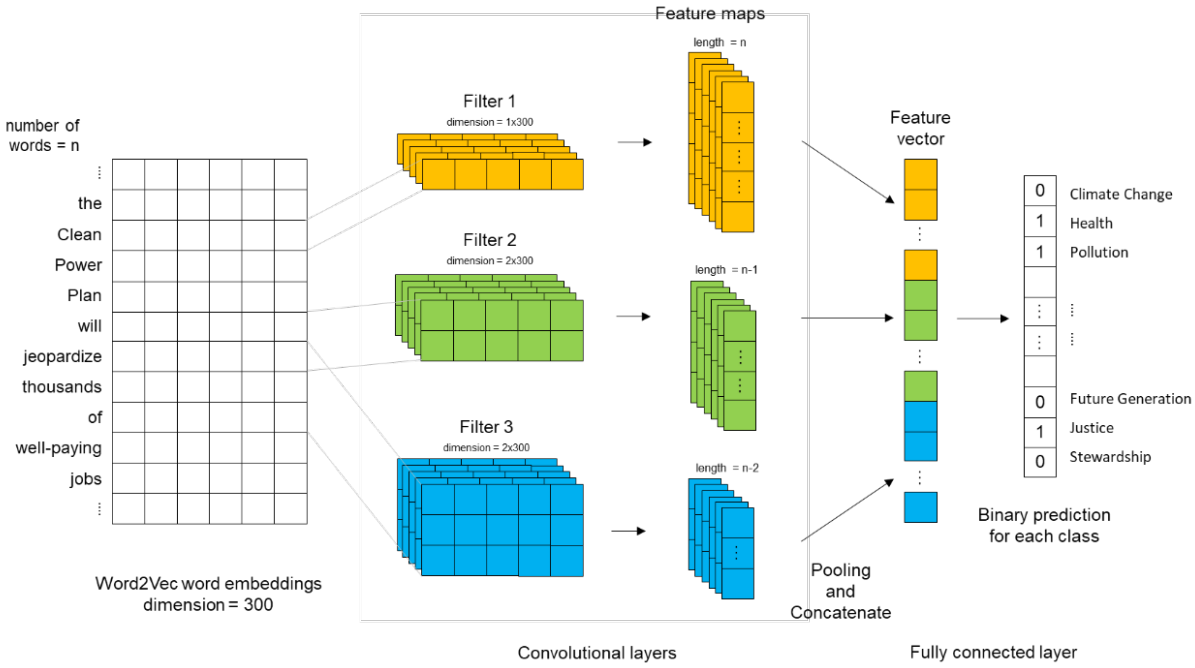


Figure 2.1 CNN model architecture for multi-label topic classification

Figure 2.1 shows the architecture of our CNN model for the multi-label topic classification. First, each comment document is translated to a *n*-by-300-dimension matrix called the word embeddings using the 300-dimension word2vec, where *n* is the number of words in each document. Word2vec embedding is a numerical representation of words, where each word is represented by a vector of numeric values. This representation is from the work of Mikolov et al. (2013), where the relationships between words were quantified based on word similarity by training very large corpus of text data. Then, this word embedding matrix goes through the convolutional layer. Here, different sizes of filters go through the word embedding matrix and creates feature maps. Then, using pooling, important information are extracted from the feature maps and they are concatenated for prediction. In fully connected layer, this concatenated feature vector is connected with the 13 topic classes, to make binary predictions for each of them. For a detailed information about CNN with word embeddings and its mechanism, see Kim (2014).

For this study, building on previous literature, we selected 1-max pooling, dropout regularization with a rate of 0.3, and a rectified linear unit (ReLU) activation function in our convolutional layer, and sigmoid activation function in the fully connected layer, as these hyper-parameters have been shown to improve accuracy (Kim 2014; Asensio et al., 2020a). In particular, the dropout technique was implemented to prevent overfitting (Srivastava et al., 2014). Other hyper-parameters include a batch size of 128; learning rate of .001; filter heights of 1, 2, and 3; 100 filters for each filter height. Filter widths are 300, which are set to the dimensionality of the word embeddings. For classifying oppose/neutral/support of the documents, only the fully connected layer is changed, where the number of target class is changed to 3 and the activation function is changed to softmax

function, as it is now a multi-class classification, where only 1 of the 3 classes should be predicted (Kalchbrenner et al., 2014, Kim 2014). Although our dataset is large for manual coding (at our pace of around 2 minutes per comment, for the full dataset of comments that are very short by qualitative data standards, labeling alone would take over 60 coding-hours per coder), it is very small in the context of training neural networks. Data augmentation is a common practice in machine learning, used to increase the amount of data available to the model, thus improving performance, by adding slightly modified copies of already existing data. Data augmentation consists of four main tasks: synonym replacement, random insertion, random swap, and random deletion (Wei and Zou, 2019). This methodology was shown to substantially increase training performance on a variety of NLP tasks, including sentiment analysis, on small datasets (Wei and Zou, 2019). In this study, we used synonym replacement using the Paraphrase Database (Ganitkevitch et al., 2013), up to 50% of the words were randomly chosen and replaced with synonym in a given comment, and followed protocols from Wei and Zou (2019) for implementation.

2.2.3.1 Multi-label topic classification

Distribution of the 200-training data for multi-label classification ranged from 65 comments assigned to natural gas (32.5%) to 155 comments to climate (77.5%), showing reasonably balanced distribution across the topics, except for stewardship topic, which was assigned to only 27 comments, accounting 13.5% of the training data. Given the general balance among the topics, data augmentation was applied to create 29 modified versions of each training comment. The train-test split ratio of 70:30 was applied to the 200 manually labeled comments, resulting in data sizes of 4,200 and 60 for training and testing data, respectively.

2.2.3.2 Multi-class sentiment classification

From the manually coded 200 comments, the distribution between oppose/neutral/support was rather highly imbalanced, with 28 comments each labeled as oppose and neutral, and 144 as support. As a multi-class classification problem, this high imbalance posed further challenges for model training – when the data were split into train and test set, less than two dozen of oppose and neutral comments were included in the train set and a handful in the test set, making it very difficult for the model to learn diverse statements for oppose and neutral classes. As a result, the authors manually coded 738 more non-random comments creating a total of 938 labeled dataset with 144 oppose, 88 neutral, and 706 support comments, resulting distribution of 15.3%, 9.4%, and 75.3% respectively. Manual coding of 738 more comments for oppose/neutral/support took the authors 4 more hours (note that manual coding took 2 minutes per comment when topics were included). Data augmentation was applied to balance the 3 classes: 5 modified versions were created for each support class comments and oppose/neutral classes were augmented with weights so the number of augmented data is similar to that of the support class. In an 80:20 train-test split, for example, if the counts of each class were 120, 70, and 560 for oppose/neutral/support classes respectively, augmented data will be $120 \cdot 5 \cdot 560//120 = 2{,}400$ for oppose class comments, $70 \cdot 5 \cdot 560//70 = 2{,}800$ for neutral class, and $5 \cdot 560 = 2{,}800$ for support class, where $//$ is the integer division operator.

2.2.4 Model performance and validation

CNN model performance is evaluated by comparing model results to manually labeled "true" results for the withheld test data described above. Table 2.2 shows the result of multi-class classification for oppose/neutral/support towards the CPP. Overall, the accuracy of the model was 86.0%, with F1 score of 0.76 with macro average. Oppose and Neutral classes have substantially lower performance than the Support class, specifically at recall. This means that there are more false negatives than false positives for these classes, likely due to the smaller representation of "Oppose" and "Neutral" in the training data, which could be overcome by oversampling in a future study. However, achieving 86.0% accuracy and 0.76 F1 score with very limited training data is quite notable.

Table 2.2 Performance of oppose/neutral/support prediction

	Accuracy (s.d.)	Precision (s.d.)	Recall (s.d.)	F1 (s.d.)
Overall	84.00 (0.03)	0.71 (0.05)†	0.73 (0.04)†	0.71 (0.04)†
Oppose	90.21 (2.38)	0.66 (0.08)	0.78 (0.09)	0.71 (0.07)
Neutral	92.84 (2.47)	0.52 (0.13)	0.53 (0.16)	0.51 (0.12)
Support	87.23 (3.14)	0.94 (0.02)	0.89 (0.05)	0.91 (0.02)

†Macro-averaged

Table 2.3 shows the overall performance results of multi-label classification of the 13 subtopics. Overall, the model achieved accuracy of 78.3% with 0.578 standard deviation. The F1 score presented in Table 2.3 is the average of the binary F1 score of each topic.

Table 2.3 Overall performance of topic predictions

Overall Performance	Value
Accuracy (%, s.d.)	78.3 (0.578)
Precision (s.d.)	0.72 (0.011)
Recall (s.d.)	0.77 (0.006)
F1 score (s.d.)	0.80 (0.006)

Table 2.4 shows accuracy and F1 score with standard deviation for each topic, based on 15 model runs with random initialization. Our CNN model achieved an F1 score above 0.80 for topics such as *Climate, Extreme Events, Coal, Future Economy, Natural Gas,* and *Clean Energy*, which we consider to be good performance, particularly given that some of these topics are highly imbalanced. *Health, Pollution, Costs,* and *Future Generations* have reasonable performance, with F1 score ranging from 0.70 to 0.76. However, *Justice* achieved a substantially lower F1 score of 0.64. No true positives were detected for *Stewardship*, leaving it without an F1 score. This failure could be due to the extreme imbalance of this topic in training data, and *Stewardship* has been removed from further analysis.

Table 2.4 Performance of topic level predictions

Topics		Accuracy (%; s.d.)		F1 (s.d.)	
Environmental Impacts	Climate	82.2	(0.87)	0.9	(0.005)
	Health	70.5	(3.02)	0.76	(0.021)
	Pollution	66.2	(2.58)	0.74	(0.017)
	Extreme Events	90	(2.19)	0.87	(0.031)
Economy	Jobs	70.1	(4.32)	0.68	(0.048)
	Costs	70.4	(3.81)	0.76	(0.026)
	Future Economy	74.8	(2.71)	0.81	(0.02)
Resources	Coal	96.9	(0.89)	0.97	(0.007)
	Natural Gas	93.8	(3.02)	0.89	(0.05)
	Clean Energy	83.7	(2.16)	0.85	(0.02)
Ethics	Future Generations	69.6	(6.44)	0.7	(0.062)
	Justice	64.6	(2.3)	0.64	(0.034)
	Stewardship	86	(0.07)	-	-

2.2.5 Analytical approach

In order to describe results from applying the trained CNN model to the entire sample of 1,900 comments (Table 2.1), we define three metrics that we use to describe public comment content across pre-implementation and pre-repeal periods ("periods") and across cities. The first metric we introduce, the *Support Score*, is an average value of comments with numeric values assigned to each class: oppose = -1, neutral = 0, support = 1, calculated as follows for each city, *i*:

$$Support\ score_i = \frac{-1 \cdot n_{oppose,\ i} + 0 \cdot\ n_{neutral,i}\ + 1 \cdot n_{support,i}}{N_i} \qquad (2.1)$$

By definition based on our trinary labeling, the *Support Score* ranges from -1 to 1, where a score of -1 represents 100% opposition and a score of 1 represents 100% support for the CPP. Due to the relatively low performance of the model in identifying opposition, we calculate the *Support Score* both for model outputs across the entire dataset (n = 1,900) and for manually labeled comments (n = 938) by city.

Our second metric, *Discussion Level*, is defined as the percentage of comments from a given subsample *i* that included a given subtopic *j*:

$$Discussion\ level_{i,j} = \frac{n_{i,j}}{N_i} \qquad (2.2)$$

Where i = city, N_i = total number of comments in city i, and $n_{oppose\,|\,neutral\,|support,\ i}$ = number of oppose/neutral/support labeled comments in city i. For example, a *Discussion Level* of 0.94 for comments from Atlanta with the *Climate* subtopic means that 94% of the comments from Atlanta addressed *Climate*.

Our final metric, *Pairwise Topic Correlation*, is defined as the pairwise correlation between subtopics within a given subsample *i*, calculated as:

$$r_i = 1 + \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} \qquad (2.3)$$

For example, a *Pairwise Topic Correlation* of 1 between *Climate* and *Justice* would mean that every comment addressing *Climate* also addresses *Justice*, and a *Pairwise Topic Correlation* of -1 would mean that the subtopics never appear together in the same comment.

2.3 Results

Figure 2.2 shows estimated *Support Score* by city for both manually labeled comments (panel a) and the model results (panel b). In all cases except for the manually labeled sample from Gillette, Support Score is positive, indicating more support for the CPP than opposition, though note that the model has a bias toward support given the highly imbalanced data. The observation that modeled Support Score is relatively high across cities other than Gillette, the city at the heart of the US' most productive coal region, is robust to both modeled and manual labels.

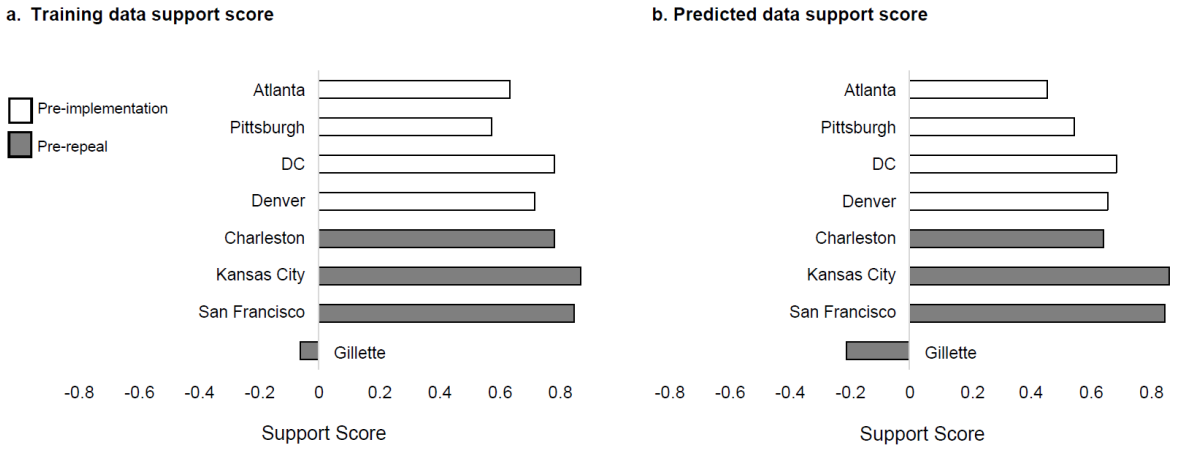


Figure 2.2 *Support Score* by city. (a) *Support Score* based on manually labeled comments (b) *Support Score* based on modeled values

Table 2.5 shows modeled *Support Score* and *Discussion Level* for 12 subtopics (excluding *Stewardship* as described above) by city. Overall, *Climate* was the most frequently discussed subtopic in all cities, with modeled *Discussion Level* between [0.89, 0.97]. The only other subtopics reaching modeled *Discussion Level* of 0.7 or higher are

Health (Charleston = 0.77, Kansas City = 0.83), *Pollution* (Charleston = 0.71, Kansas City = 0.75), *Costs* (Atlanta = 0.70, Pittsburgh = 0.75), *Future Economy* (Pittsburgh = 0.80, Gillette = 0.77), and *Coal* (Gillette = 0.74).

Table 2.5 Modeled *Support Score* and *Discussion Level* by city

		Pre-Implementation				Pre-Repeal			
		Atlanta	Pittsburgh	DC	Denver	Charleston	Kansas City	San Francisco	Gillette
	Support Score	0.46	0.58	0.69	0.66	0.64	0.86	0.85	-0.21
Environmental Impacts	Climate	0.94	0.97	0.96	0.96	0.89	0.93	0.92	0.89
	Health	0.54	0.63	0.64	0.60	0.77	0.83	0.69	0.52
	Pollution	0.53	0.60	0.58	0.62	0.71	0.75	0.59	0.59
	Extreme Events	0.18	0.25	0.34	0.33	0.29	0.44	0.37	0.22
Economy	Jobs	0.36	0.60	0.39	0.46	0.54	0.54	0.29	0.66
	Costs	0.70	0.75	0.66	0.67	0.63	0.56	0.42	0.69
	Future Economy	0.65	0.80	0.64	0.69	0.66	0.64	0.51	0.77
Resources	Coal	0.53	0.64	0.44	0.55	0.67	0.49	0.23	0.74
	Natural Gas	0.28	0.27	0.23	0.28	0.23	0.10	0.06	0.25
	Clean Energy	0.44	0.55	0.53	0.55	0.46	0.49	0.39	0.35
Ethics	Future Generations	0.39	0.44	0.46	0.46	0.54	0.60	0.53	0.32
	Justice	0.36	0.45	0.40	0.29	0.61	0.61	0.33	0.49

Overall, *Climate* was the most frequently discussed subtopic in all cities, with modeled *Discussion Level* between [0.89, 0.97]. The only other subtopics reaching modeled *Discussion Level* of 0.7 or higher are *Health* (Charleston = 0.77, Kansas City = 0.83), *Pollution* (Charleston = 0.71, Kansas City = 0.75), *Costs* (Atlanta = 0.70, Pittsburgh = 0.75), *Future Economy* (Pittsburgh = 0.80, Gillette = 0.77), and *Coal* (Gillette = 0.74).

Figure 2.3 compares aggregated *Discussion Level* for pre-implementation and pre-repeal hearings and listening sessions by subtopic for both modeled and labeled results,

with statistically significant differences at the a= 0.05, 0.01, and 0.001 levels as determined by a two-sample t-test marked.

Figure 2.3 Comparison of *Discussion Level* between pre-implementation and pre-repeal phases.

Eight out of 12 topics had statistically significant differences for model results between the two periods at a significance level of $a = 0.05$. Although interpretive caution is advised for modeled results, similar findings hold for the subset of manually labeled comments.

Figure 2.4 shows the results for pairwise topic discussion level correlation. First, we see that support score strongly correlated with Environmental Impacts topics, Clean Energy, and Future Generation mentions. For Environmental Impacts theme, we see that the topics are negatively correlated with Economy, positively with Clean Energy and Ethics. Economy and fossil fuel-related topics are strongly correlated. In Ethics theme, Justice is

strongly correlated, positively with Health and Pollution, as well as Jobs. Interestingly, Justice has practically no correlation with support score, implying that justice is emphasized for both supporting and opposing attitudes towards the plan. More on this result will be discussed in next section.

		1	2	3	4	5	6	7	8	9	10	11	12	13
	1. Support Score	1												
Environmental Impacts	2. Climate	0.42	1											
	3. Health	0.70	-0.16	1										
	4. Pollution	0.37	-0.32	0.87	1									
	5. Extreme Events	0.70	0.09	0.78	0.64	1								
Economy	6. Jobs	-0.56	-0.27	0.00	0.39	-0.21	1							
	7. Costs	-0.51	0.32	-0.53	-0.28	-0.68	0.54	1						
	8. Future Economy	-0.62	0.14	-0.41	-0.06	-0.53	0.84	0.87	1					
Resources	9. Coal	-0.68	-0.26	-0.27	0.14	-0.57	0.86	0.78	0.88	1				
	10. Natural Gas	-0.51	0.26	-0.65	-0.41	-0.75	0.34	0.93	0.73	0.71	1			
	11. Clean Energy	0.54	0.81	0.21	0.17	0.26	0.01	0.39	0.24	0.02	0.32	1		
Ethics	12. Future Generations	0.86	0.00	0.96	0.76	0.84	-0.24	-0.63	-0.57	-0.48	-0.68	0.30	1	
	13. Justice	-0.06	-0.53	0.62	0.78	0.19	0.63	0.03	0.22	0.46	-0.18	-0.11	0.37	1

Figure 2.4 Pairwise topic discussion level correlation

2.4 Discussion

2.4.1 Topic discussion levels

The topic discussion level comparison between the pre-implementation and pre-repeal periods shown from the results provides important implications on the dynamics of topics that are associated with the sentiment towards the policy.

Health and pollution topics discussion level has significantly increased in the pre-repeal comments. The reason for this may be because speakers not wanting to lose the co-benefits of climate policy focused on closing coal plants. This is in line with the pairwise topic discussion level correlation results, where these two topics are positively correlated with the support score. On the other hand, discussion level on Jobs increased significantly, and it showed high negative correlation with the support score (-0.56). Given the very high positive correlation with Coal (0.86), this implies that commenters opposed to the plan or supported the repeal, often discussing job losses of coal industry workers. This may also explain why Natural Gas and Clean Energy discussion levels have decreased in the repeal period, as the focus on resources was put on coal instead of others, and jobs in the coal industry.

An important finding here is the Justice topic. We saw a big increase in the repeal period, and interestingly Justice topic has almost no correlation with the support score (-0.06). This can be interpreted as there is bifurcated topicality with environmental justice versus states' right and just transition contents. People arguing on "justice for the coal workers" versus arguing fairness to the future generations on protecting the environment. For example, comments saying *"...the plan is doing injustice for coal mine workers..."* or *"... it is only fair for the future generations that we protect the environment..."*, show the bifurcated discussion on Justice topic, respectively.

2.4.2 Political environment for commenters

One of the most striking observations from this descriptive research is that *Support Score* is consistently high, indicating more support than opposition, in all cases other than

for Gillette, Wyoming. Even for Gillette, *Support Score* is not as highly negative as it is positive in other areas. This is notable in part because, as noted above, Gillette is the core community in the country's most productive coal mining region, the Powder River Basin. Further, although this is not necessarily clear from available text data, the listening session in Gillette was in the same location, and immediately after, a pro-coal professional function hosted by the Wyoming Mining Association, National Mining Association, and Friends of Coal. Based on one author's (Grubert's) observations as an attendee, those who attended the pro-coal event were relatively numerous and more formally dressed (e.g., in suits) than others. Further reinforcing the observation that the pro-coal event was a professional and attractive event, several particularly high-ranking speakers (including the governor of Wyoming, both the state's US Senators, and several industry representatives) were placed at the very beginning of the speaker schedule (see Figure 2.5), with one effect that the first pro-CPP comment was well into the first session (and received audibly less applause than the extensive pro-repeal comments). Pro-CPP comments in Gillette should be interpreted as taking place in the context of a highly public setting in a relatively small community.

Given the preponderance of CPP support in other sessions, similar public pressure considerations could apply in a manner that might have reduced oppositional statements rather than supportive statements. The only other session one of us attended was in San Francisco, a much larger city than Gillette, with the additional comment that the San Francisco session was divided into multiple rooms, less heavily inflected by the presence of local power, and generally organized such that speakers were not as publicly performing to an audience.

Figure 2.5 Coal industry-supported "Coffee and Conversation" event board; photographs taken by E. Grubert outside the Gillette, Wyoming Clean Power Plan Listening Session, 27 March 2018

2.4.3 Limitations

This analysis has several limitations, due both to typical challenges associated with machine learning and to specific challenges associated with this research topic. CNN model results should be interpreted with caution, particularly relative to manually coded data, due to performance constraints and a limited training set as described above. The modeling concern we expect has the largest qualitative impact on our findings is the relatively low F1 score for detecting comments neutral to the CPP, which is partly a result of having limited opposed comments in the training set. Nonetheless, based on spot checks, document

review, and other CPP context, the idea that most in-person comments supported the CPP seems valid.

Relatedly, the machine learning task here is classification rather than interpretation, and our thematic binning (and subsequent use of counts and proportions to describe content data) does not capture important interpretive nuances about how speakers were incorporating subtopics into their comments. For example, speakers claiming that climate change is a hoax or an existential threat would both result in a *Climate* label. Our use of manual coding as a training input to a CNN also presented a challenge in that we as coders knew that our labels would be used to train a model: although we attempted to avoid adapting our codes in ways that would make the CNN task easier but might not be best-practice for manual coding, such adaptations might have occurred, particularly for broad subtopics that lend themselves easily to keyword-based coding. The use of machine learning for this particular task also poses limitations, as with identifying important but rare or complex themes like *Stewardship* (Hazboun et al., 2019). The use of synonym replacement and other techniques intended to improve CNN performance could mask specific contextual meanings associated with certain terms, like "just transition" or "clean energy."

One major specific challenge with this research is that the issue under consideration for the pre-implementation and pre-repeal comments was not directly inverted. Most relevantly, some commenters opposed to implementation of the CPP were opposed because they wanted a stronger climate policy, and others because they wanted no climate policy. This diversity in meaning of "oppose" could have contributed to modeling challenges. Further, evaluating public comments is not the same as evaluating public opinion, given

that commenters are not a random sample of the relevant population. In this case, particularly given limited access to in-person hearings or listening sessions in some cases, city-based comments are also not restricted to residents of the region, so regional trends should be interpreted with care. For example, the Charleston, WV hearing was initially planned as the only session associated with the proposed repeal and thus attracted many non-local residents, which could partly explain the large difference between attitudes at the Gillette and Charleston sessions despite both being coal mining areas. Finally, we did not attend all meetings in person. As described above, contextual information about side events and meeting layout can be important for interpretability.

2.5 Conclusion

In this chapter, we achieved overall 0.71 and 0.80 F1 scores for multi-class sentiment and multi-label topic classifications on highly targeted public comments data. Given the lack of labeled data in this particular dataset, we have successfully overcome limited training data challenges using NLP with deep learning, with data augmentation techniques. This study provides a potential step change in our ability to aggregate data and insights for large-scale public comments data to support efficient policymaking process.

Taking advantage of the computational advance, we find that although the CPP was eventually repealed, most of the spoken comments were supportive. Health and pollution issues were significantly more discussed during the pre-repeal period, correlated highly with the supporting comment. Justice discussion largely increased during pre-repeal,

showing bifurcation between environmental justice (health and pollution), versus states' right and just transition content (jobs and coal).

Further improvements on this study are two folds. First, given the limited number of labeled data, it may be useful to implement different model, particularly transfer learning, where we can benefit from pre-trained models such as the bidirectional encoder representations (Devlin et al., 2019) and the universal language model fine-tuning (Howard and Ruder, 2018), that has linguistic knowledge prior to training with our dataset. However, given that these models introduce complexity and require much larger computational resources, CNN can still be a more feasible algorithm for social scientists.

Second, the public comment collection has more than 3 million submitted comments for the CPP in the database. Using the models trained in this study, large-scale analysis of these data will lead to important findings on the CPP.

2.6 Acknowledgements

This research was supported by the Russell Sage Foundation "Small Grants in Computational Social Science" program (grant G-1903-13611).

CHAPTER 3. TOPIC CLASSIFICATION OF ELECTRIC VEHICLE CONSUMER EXPERIENCES WITH TRANSFORMER-BASED DEEP LEARNING[2]

3.1 Introduction

In recent years, there has been a growing emphasis on vehicle electrification as a means to mitigate the effects of greenhouse gas emissions (EPA 2018a) and related health impacts from the transportation sector (NRC 2010). For example, typical calculations suggest that electric vehicles reduce emissions from 244 to 98g/km, and this number could further decrease to 10g/km with renewable energy integration (Hoekstra 2019). The environmental benefits range by fuel type with reported carbon intensities of 8,887 grams CO_2 per gallon of gasoline, and 10,180 grams CO_2 per gallon of diesel EPA (2018b). Government-driven incentives for switching to electric vehicles, including utility rebates, tax credits, exemptions and other policies, have been rolled out in many U.S. states (DOE 2019; Carley et al., 2013; Sheldon et al., 2017). In this effort, public charging infrastructure remains a critical complementary asset to consumers in building range confidence for trip planning and in EV purchase decisions (Hardman et al., 2018; Anderson et al., 2018; Brückmann and Bernauer 2020). Prior behavioral research has shown that policies designed to enhance EV adoption have largely focused on increasing the quantity of cars

[2] This chapter was published as a journal article in the Patterns journal by the Cell Press with Daniel J. Marchetto, Sameer Dharur, and Omar I. Asensio as the co-authors. The citation for the journal article is as follows: Ha, S., Marchetto, D. J., Dharur, S., & Asensio, O. I. (2021). Topic classification of electric vehicle consumer experiences with transformer-based deep learning. Patterns, 2(2), 100195. https://doi.org/10.1016/j.patter.2020.100195

and connected infrastructure as opposed to the quality of the charging experience (Asensio et al., 2020a). However, a fundamental challenge to deploying large-scale EV infrastructure is regular assessments of quality.

Private digital platforms such as mobility apps for locating charging stations and other services have become increasingly popular. Reports by third party platform owners suggest there are already over 3 million user reviews of EV charging stations in the public domain (Recargo 2020; Chargemap 2020; Open Charge Map 2020). In this paper, we evaluate whether transformer-based deep learning models can automatically discover experiences about EV charging behavior from unstructured data and whether supervised deep learning models perform better than human benchmarks, particularly in complex technology areas. Because mobile apps facilitate exchanges of user texts on the platform, multiple topics of discussion exist in EV charging reviews. For example, a review states: *"Fast charger working fine. Don't mind the $7 to charge, do mind the over-the-phone 10 minute credit card transaction."* A multi-label classification algorithm may be able to discover that the station is functional, that a user reports an acceptable cost, and that a user reports issues with customer service. Therefore, text classification algorithms that can automatically perform multi-label classification are needed to interpret the data. Being able to do multi-label classification on these reviews is important for three principal reasons. First, these algorithms can enable analysis of massive digital data. This is important because behavioral evidence about charging experiences has primarily been inferred through data from government surveys or simulations. These survey-based approaches have major limitations as they are often slow and costly to collect, are limited to regional sampling, and are often subject to self-report or recency bias. Second, multi-label

algorithms with digital data can characterize phenomena across different EV networks and regions. Some industry analysts have criticized EV mobility data for poor network interoperability, which prevents data from easily being accessed, shared and collected (Recharge 2020). This type of multi-labeled output is also important for application programming interface (API) standardization across the industry such as with emerging but not yet widely accepted technology standards including the Open Charge Point Protocol (Open Charge Alliance, 2020a) that would help with real-time data sharing across regions. Third, this capability may be critical for standardizing software and mobile app development in future stages of data science maturity (see https://www.cell.com/patterns/dsml) to detect behavioral failures in near real-time from user generated data.

Modern computational algorithms from natural language processing (NLP) could uniquely address the need for fast, real-time consumer intelligence related to electric mobility, but these algorithms need to be appropriately tailored to domains to be useful. Large-scale analysis of unstructured EV user data remains difficult to carry out, especially when there are multiple topics discussed in each review, and the datasets are imbalanced. Unbalanced data creates challenges for models to learn important but less frequently occurring labels often lead to algorithmic bias. In this paper, we demonstrate the use of deep neural networks to automatically discover insights for topic analysis. We use supervised learning to overcome prior challenges with unsupervised methods that could produce clusters with very little theoretical or social meaning. We provide a proof of concept to the complex task of multi-label topic classification in this domain, which builds on an earlier demonstration of binary sentiment classification with NLP (Recargo 2020).

We apply transformer neural networks, a recent class of pre-trained contextual language models, to accurately detect long-tail discussion topics with imbalanced data—a capability that has been elusive with prior approaches.

Prior research demonstrated the efficacy of convolutional neural networks (CNNs; LeCun and Bengio 1998; Kim 2014; Zhang and Wallace 2017; Yin et al., 2017) and long short-term memory (LSTM), a commonly used variant of recurrent neural networks (RNNs; Yin et al., 2017; Hochreiter and Schmidhuber 1997) for NLP. These models have been recently applied to sentiment classification and single-label topic classification tasks in this domain. As a result, this has increased our understanding of potential EV charging infrastructure issues such as the prevalence of negative consumer experiences in urban locations as compared to non-urban locations (Asensio et al., 2020a; Alvarez et al., 2019; Ha et al., 2020). While these models showed promise for binary classification of short texts, generalizing these models to reliably identify multiple discussion topics automatically from text presents researchers with an unsolved challenge of under-detection, particularly in corpora with wide-ranging topics and possible imbalances in the training data. Prior research using sentiment analysis indicates negative user experiences in EV charging station reviews, but it has not been able to extract the specific causes (Asensio et al., 2020a). As a result, multi-label topic classification is needed to understand behavioral foundations of user interactions in electric mobility.

In this paper, we achieve state-of-the-art multi-label topic classification in this domain using transformer-based (Vaswani et al., 2017) deep neural networks BERT, which stands for bidirectional encoder representations (Devlin et al., 2019) and XLNet, which integrates ideas from Transformer-XL (Yang et al., 2019) architectures. We benchmark the

performance of these transformer models against classification results obtained from adapted CNNs and LSTMs. We also evaluate the potential for super-human performance of the classifiers by comparing human benchmarks from crowd annotated training data, versus expert annotated training data and transformer models. The extent of this improvement could significantly accelerate automated research evaluation using large-scale consumer data for performance assessment and regional policy analysis. We discuss implications for scalable deployment, real-time detection of failures, and management of infrastructure in sustainable transportation systems.

3.2 Methods

3.2.1 Data

We reanalyze data derived from a nationally representative collection of unstructured consumer reviews from 12,720 charging station locations across the United States. It comprises 127,257 reviews all written in English by 29,532 registered and unregistered EV drivers across a 4-year duration from 2011 to 2015 (Asensio et al., 2020a; Alvarez et al, 2019; Asensio et al., 2020b).

The spatial coverage of the dataset includes reviews from 750 metropolitan statistical areas (309 large MSAs of population 1 million or more; 228 medium MSAs population of 250,000-999,999; 213 small MSAs population of 50,000-249,999). This also includes 294 micropolitan statistical areas (e.g. μSA population 10,000-49,999), and 232 non-core-based statistical areas (e.g. population less than 10,000). This spatial coverage is

based on the 2013 OMB delineation of metropolitan statistical areas (MSA) and micropolitan statistical areas.

The data is statistically representative of the entire U.S. EV market, which includes all major EV networks, and a mix of both public and private stations, urban and rural stations, and both low and highly rated stations. The data includes the text of consumer reviews and contains other useful indicators such as the timestamp of the reviews, the car make and model. We also geo-coded the station location and related points of interest using the Google Places API. However, the dataset does not contain EV transactions data, such has how many kWh were transferred. The data is also only observable conditional on a user checking-in and posting a review.

This type of data is expanding globally and we estimate that there are already over 3.2 million reviews through 2020 across more than 15 charge station locator apps (Regarco 2020; Chargemap 2020; Open Charge Map 2020; ChargePoint 2020; Recharge 2020). This includes English-language reviews as well as reviews in over 42 languages in all continents, such as Ukrainian, Russian, Spanish, French, German, Finnish, Italian, Croatian, Icelandic, Haitian-creole, Ganda, Sudanese, Kinyarwanda, Afrikaans, Nyanja, Korean, Mandarin, Japanese, Indonesian and Cebuano.

3.2.2 Developing the Coding Scheme for Supervised Learning

We developed the coding scheme for our typology from prior work and theory using three strategies. First, we reviewed the extant literature to capture the most important potential behavioral issues for EV drivers. This led to identification of Range Anxiety (Carley et al., 2013; Rauh et al., 2015; Jung et al., 2015; Noel and Sovacool 2016; Egbue

and Long 2012), Dealership practices (Rubens et al., 2018; Lynes 2018), Cost (Carley et al., 2013; Egbue and Long 2012; Hidrue et al., 2011; Nicolson et al., 2017; Kühl et al., 2019), Service Time (Carley et al., 2013; Egbue and Long 2012; Hidrue et al., 2011; Kühl et al., 2019), Availability issues (Kempton et al., 2001; Liao et al., 2017), User Interaction (Burgess et al., 2013; Morstyn et al., 2018; Lee et al., 2020), station Functionality (Asensio et al., 2020a; Kühl et al., 2019; NRC 2015), and Location (Asensio et al., 2020a). Second, to find evidence of the importance of these topics from the data, we hand-coded 8,953 randomly selected reviews to validate the 8 topics from prior literature and used these to generate 34 sub-topics for classification. We found that only 1% of the reviews were unclassifiable according to our 8 main categories (e.g. Other). Third, to validate the coding scheme, we also interviewed industry experts and practitioners, which allowed us to further refine our main topics and sub-topics shown in Table 3.1. This included representatives from firms such as General Motors, Chargepoint, Recharge Technologies, Electrada, Electrify America, and charging station managers (e.g. representatives from Ford and Georgia Tech Parking and Transportation Services).

Table 3.1 EV mobile app typology of user reviews (Ha et al., 2021)

Topic	Subtopic Examples
Functionality	General Functionality, Charger, Screen, Power Level, Connector Type, Card, Reader, Connection, Time, Error Message, Station, Mobile Application, Customer Service
Range Anxiety	Trip, Range, Location Accessibility
Availability	# of Stations Available, ICE, General Congestion
Cost	Parking, Charging, Payment
User Interactions	Charger Etiquette, Anticipated Time Available, User Tips
Location	General Location, Directions, Staff, Amenities, Points of Interest, User Activity, Signage
Service Time	Charging Rate
Dealership	Dealership Charging Experience, Competing Brand Quality, Relationship with Dealers
Other	General Experiences

3.2.3 Human Annotation of Training Data

A common criticism with deep neural networks is the high cost and annotator skill requirements for implementations in specialized corpora. We evaluated possible methods to lower implementation costs, such as crowd sourcing by using online labor pools for human annotation. This led us to conduct human annotator experiments with two training sets each labeled by a crowd of non-experts and a small group of trained experts. Given the known possible biases with historical data, we investigated whether protocols related to the labeling of the training data could have an impact on performance (Rambachan et al., 2020; Cowgill and Tucker 2017).

The crowd and expert annotators each labeled a random sample of 10,652 reviews. We used an 80:10:10 split for training, validation, and testing, which met our objective of

having an equal number of training data for both annotator groups. We conducted statistical tests to determine whether the sampled training dataset is representative of the full dataset in key observable station characteristics. We confirmed that the training dataset is statistically representative in the mix of urban and non-urban stations (t-test p-value 0.426), public and private stations (t-test p-value 0.709), as well as by station points of interest (t-test p-value 0.802), e.g. retail, shopping, workplace, and transit centers, etc.). We also found that the training data was not statistically different in topic distribution from the predictions of the full dataset (Kolmogorov-Smirnov test p-value 0.9801).

3.2.3.1 Crowd Annotators.

For the crowd-sourced training data sample, 1,000 U.S. adults (age 18+) were pre-recruited via a Qualtrics online panel using their popular online survey platform. The crowd was statistically sampled on the basis of age, income, education, and sex, representative of the U.S. population. This is important to mitigate possible human rater biases that could arise when discussing environmental topics. To enhance understanding of the domain-specific terminology for the general crowd, definitions and examples for the topics and sub-topic as shown in Table 3.1 were provided for annotation along with a supporting diagram containing typical components of an EV charging station (See Figure 3.1 and Figure 3.2). We report the Fleiss' Kappa for crowd annotators as 0.007.

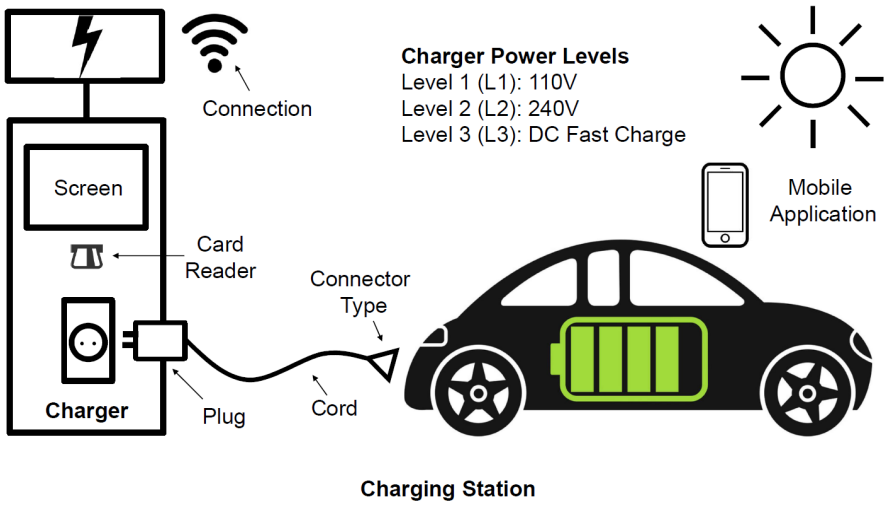


Figure 3.1 Diagram of EV charging station (Ha et al., 2021)

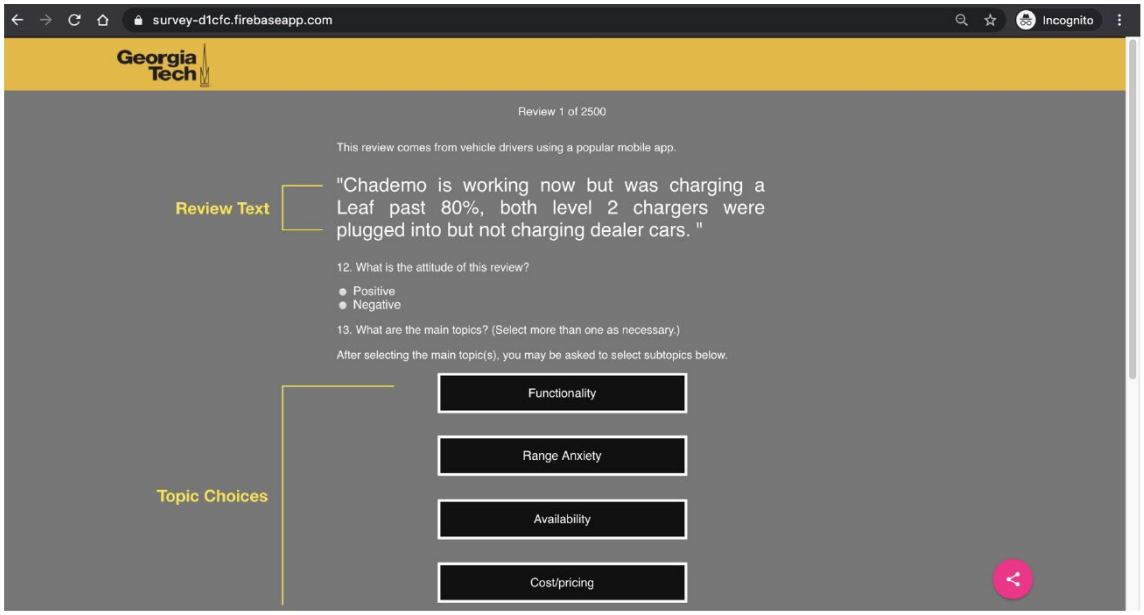


Figure 3.2 Web app for training data collection (Ha et al., 2021)

3.2.3.2 Expert Annotators.

For the expert-sourced training data sample, five student annotators with technical backgrounds were recruited and trained in a facilitated focus group. They were instructed to recognize the domain-specific topics using a detailed training manual for the annotation. To support scientific replication and to document the protocols, we have open sourced this training manual (Appendix B). These protocols were developed in consultation with EV industry experts who have been in contact with the researchers. Although our expert annotators have been trained to recognize domain-specific terminology, we acknowledge that we are not able to compare the performance of our expert annotators to EV industry professionals due to cost reasons. Despite this limitation however, we find that our human experts are two orders of magnitude more reliable in the annotation (76-fold increase in our reliability measure) versus the crowd annotators (κ= 0.538 and κ= 0.007, respectively). See the Model Metrics section under Performance Measures for additional details on computing Fleiss' Kappa.

To provide a greater control over the labeling task, we developed a custom web application used by the expert annotators as shown in Figure 3.2. The web app provides efficient database support for random sampling from a large dataset and overcomes latency and scaling challenges that we encountered during crowd annotation in popular survey software.

3.2.3.3 Ground Truth Labels.

To generate the ground truth labels, we followed the same training protocols used by the expert annotators. Then, we randomly sampled 100 overlapping reviews that were

annotated by both annotator groups to enable performance comparisons. On this sample, we conducted an additional round of researcher audits that validated 100% agreement on the annotations. Given that the human experts exhibited some level of disagreement (Fleiss' kappa = 0.538), this sample was used to benchmark the performance of the U.S. crowd and the human experts. To generate the uncertainty, we performed a cross validation using block randomization with 10 equal-sized blocks of ground truth data.

3.2.4 Performance Measures

3.2.4.1 Model Metrics.

In order to assess model performance, we report the macro-averaging F1 score, which is a standard metric for classifier performance on detection of false positives and false negatives. We use standard measures for multi-label accuracy, where annotators could choose multiple labels per review. Our overall accuracy metric accounts for partially correct matches. By convention, this is equivalent to 1 - Hamming Loss, where the Hamming Loss is an xor calculation of the dissimilarity (i.e. a fraction of wrong labels compared to the total number of labels). For L categories classified on a sample of size N, the accuracy can be calculated as:

$$\begin{aligned} \text{Overall Accuracy} &= 1 - \text{Hamming Loss} \\ &= 1 - \frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} xor(y_{i,j}, z_{i,j}) \end{aligned} \tag{3.1}$$

For example, if a multi-label prediction [1, 1, 1, 0] had a true label [1, 1, 1, 1], the accuracy is 3/4 or 75%.

3.2.4.2 Inter-Rater Reliability.

To measure the inter-rater agreement level among the annotators, we used Fleiss' Kappa (κ), which allows for the measurement of agreement between multiple annotators (e.g., more than 2). It is calculated as below:

$$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}, \tag{3.2}$$

where $\bar{P}$ is the average number of agreements on all annotations between rater pairs for the reviews, and $\bar{P_e}$ is the sum of squares of the probability share for the assignment to a topic. As κ is bounded between -1 and 1, when κ is less than 0, agreement between raters is occurring below what would be expected at random, while a κ above 0 means that agreement between raters is occurring more than what would be expected by random chance (Landis and Koch 1977). For more information, see Fleiss (1971).

3.2.5 Ethics Statement

Human subjects research was conducted under the approved Institutional Review Board (IRB) Protocol No. H18250.

3.3 Results & Discussion

3.3.1 Discovering Topics

Charging station reviews can be considered asynchronous social interactions within a community of EV drivers. To characterize user experiences, we introduce 8 main topics and 32 sub-topics that make up a typology of charging behavior. This typology allows for easier identification of behavioral issues with the charging process (Table 3.1). The definitions we use for supervised learning are as follows: *Functionality* refers to comments describing whether particular features or services are working properly at a charging station. *Range Anxiety* refers to comments regarding EV drivers' fear of running out of fuel mid-trip and to comments concerning tactics to avoid running out of fuel. *Availability* refers to comments concerning whether charging stations are available for use at a given location. *Cost* refers to comments about the amount of money required to park and/or charge at particular locations. *User Interaction* refers to comments in which users are directly interacting with other EV drivers in the community. *Location* refers to comments about various features or amenities specific to a charging station location. The *Service Time* topic refers to comments reporting charging rates (e.g. 10 miles of range per hour charged) experienced in a charging session. The *Dealerships* topic refers to comments concerning specific dealerships and user's associated charging experiences. Reviews that do not fall into the previous 8 topics refer to the *Other* topic, which are relatively rare.

In preliminary experiments, we investigated several unsupervised topic modeling techniques that did not provide theoretically meaningful clusters. By contrast, our empirically driven typology is ideally suited for hypothesis testing, spatial analysis,

benchmarking with other corpora in this domain, and real-time tracking of station failures, all of which are not identifiable with current information systems. For additional details on how the typology and coding scheme were developed from prior work and theory, see Developing the Coding Scheme for Supervised Learning section.

3.3.2 Transformers Beat Other Deep Neural Networks

3.3.2.1 Overall Performance.

We evaluated the accuracy of BERT and XLNet transformer models against other leading models, CNN and LSTM, which were previously dominant architectures in this domain (Asensio et al., 2020a; Ha et al., 2020). Given that we have imbalanced data for machine classification, we also report the F1 score, which is the harmonic average of precision and recall, and is considered a measure of detection efficiency. As shown in Table 3.2, we achieved high overall accuracy scores for BERT and XLNet of 91.6% (0.13 s.d.) and 91.6% (0.07 s.d.), and F1 scores of 0.83 (0.0037 s.d.) and 0.84 (0.0015 s.d.), respectively. The standard deviations were generated from 10 cross-validation runs. While CNN and LSTM models had slightly lower accuracy, we find that both transformer models outperform the CNN and LSTM models considering both accuracy and F1 score. We report 2 to 4 percentage point improvements in the F1 scores for both transformer models. For reference, we provide the hyper-parameters used for the transformer models in Table 3.3.

Table 3.2 Overall model performance (Ha et al., 2021)

	Accuracy % (s.d.)	F1 score (s.d.)
BERT	91.6 (0.13)	0.83 (0.0037)
XLNet	91.6 (0.07)	0.84 (0.0015)
Majority Classifier	81.1 (0.00)	0.45 (0.0000)
LSTM	90.3 (0.17)	0.80 (0.0036)
CNN	90.9 (0.12)	0.81 (0.0032)

Note: Models are trained and tested on expert annotated data

Table 3.3 Hyper-parameters for BERT and XLNet (Ha et al., 2021)

Hyper-parameter	Value
Number of Epochs	20
Batch Size	8
Learning Rate	1.00E-04
Max Sequence Length	512
Weight Decay	0.01
Adam Epsilon	1.00E-08
Max Grad Norm	1
Warmup Steps	500
Train:Valid:Test	80:10:10

The F1 scores for the transformer models are also a substantial 40 percentage points higher compared with the majority classifier (Table 3.2). This means the models learned to detect minority classes effectively. Briefly, the majority classifier provides a measure of the level of imbalance. For a given category, the majority classifier simply predicts the most prevalent label. For example, if 90% of training data has not been selected for a topic, then the classifier predicts all data as not selected, giving a high accuracy of 90%. Thus, for highly imbalanced data, a majority classifier can provide arbitrarily high accuracy without significant learning (Schapire 1990). Because it is possible that mis-classification errors may not distribute equally across the topics, in the next section, we also evaluated the performance by topics.

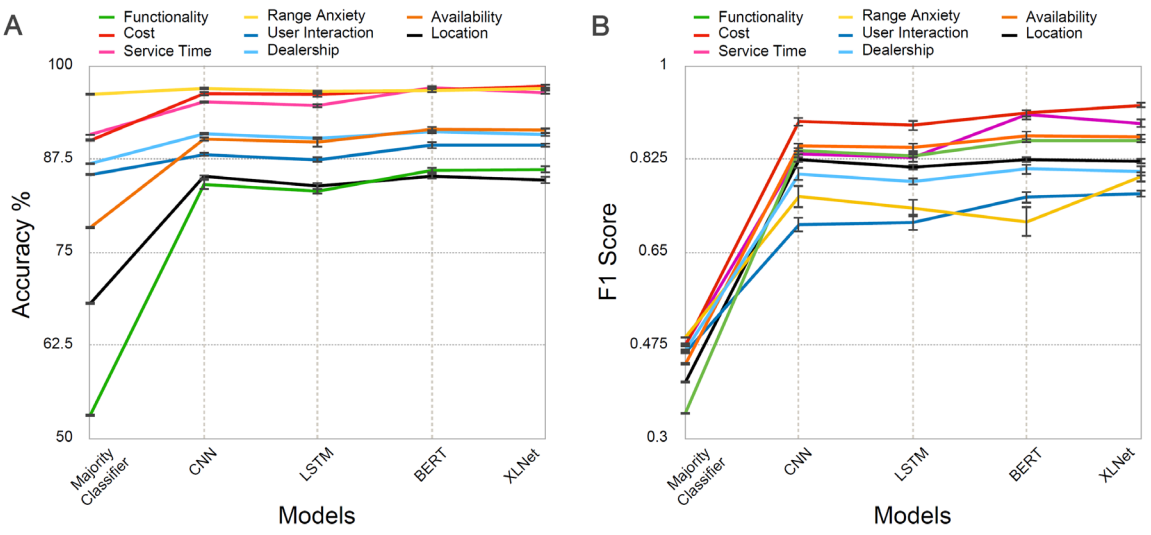


Figure 3.3 Topic level classification performance. (a) Accuracy and (b) F1 score (Ha et al., 2021)

3.3.2.2 Increasing Detection of Imbalanced Labels.

A key challenge was to evaluate whether we could improve multi-label classifications even in the presence of imbalanced data. Figure 3.3 (A) shows a large

percentage point increase in accuracy for all the deep learning models tested, as compared with the majority classifier. This evidence of learning is especially notable for the most balanced topics (e.g. *Functionality, Location and Availability*). As shown in Figure 3.3 (B), we report improvements in the F1 scores for BERT and XLNet across most topics versus the benchmark models. In particular, this result holds for the relatively imbalanced topics (e.g. *Range Anxiety, Service Time*, and *Cost*), which have presented technical hurdles in prior implementations (Ha et al., 2020). In comparison with the previously leading CNN algorithm, BERT and XLNet produce F1 score increases of 1-3 percentage points on *Functionality, Availability, Cost, Location*, and *Dealership* topics, 5-7 percentage points on *User Interaction*, and *Service Time* topics. For *Range Anxiety*, BERT is within the statistical uncertainty of the CNN performance, while XLNet produces an increase in the F1 score of 4 percentage points. These numbers represent considerable improvements in topic level detection.

Given these promising results, next we consider some requirements for possible large-scale implementation related to computation time and scalability related to the sourcing of the training data.

3.3.3 Computation Time

An important metric to consider while running deep learning models for large-scale deployment is the computation time. Deep neural networks have been criticized for the large amount of resources needed such as graphics processing units (GPUs) and distributed computing clusters, frequently leading to higher costs of deployment (Yan et al., 2015). Further, NLP researchers have also considered the environmental costs of the power

consumption and CO_2 emissions for computing (Strubell et al., 2019), which necessarily involve trade-offs. In our application, we report the training times per epoch for BERT and XLNet as 196 and 346 seconds, respectively. These results were generated using 4 widely available NVIDIA Tesla P100 GPUs with 16 GB of memory.

We find that the training and testing times are considerably longer for the transformer models compared with CNN and LSTM. For transformers, total computing times vary from 1 to 4 hours and for CNN and LSTM, computing times vary from 1 to 90 minutes, depending on the number of GPUs. We argue that the model performance improvements in the transformer models may be justified for large-scale deployment. This is because the increase in computational cost is offset by substantial gains in accuracy and F1 score. When comparing BERT and XLNet within the class of transformers, we also show BERT to be considerably faster in total computing time for a comparable level of performance. Therefore, we note that as further enhancements to BERT and its optimized variants are rapidly advancing in the literature (Lan et al., 2020; Liu et al. 2020; Sanh et al., 2019), we argue that BERT could be a preferred text classification algorithm for this domain. In the next section, we consider scalability of the models by evaluating potential sources of training data.

3.3.4 Trained Experts Beat the Crowd

In Table 3.4, we compare the machine classification results based on training data from a crowd of non-experts versus a group of trained expert annotators. For performance comparison of models trained with expert and crowd annotated data, we created a ground truth dataset by conducting researcher audits to ensure 100% agreement on the ground truth

labels. See Human Annotation of Training Data section for further details. Not surprisingly, we find that human experts are closer to the ground truth (random holdout sample; $n = 100$) in both accuracy and F1 score as shown in Table 3.4. This is consistent with related literature on limitations to wise crowds (Surowiecki 2005). In fact, prior research has found gaps in general public knowledge about EVs and consumer misperceptions (Roberson and Helveston 2020; Krause et al., 2013; Axsen et al., 2017; Wang et al., 2018). In the next section, we quantify the performance of crowd-trained versus expert-trained transformer models, using the two experimentally curated sources of training data.

Table 3.4 Ground truth evaluation of human performance versus transformer models (Ha et al., 2021)

Classifier	Training set	Accuracy % (s.d.)	F1 score (s.d.)
BERT	Expert-annotated	89.1 (4.09)	0.82 (0.06)
BERT	Crowd-annotated	73.2 (3.85)	0.53 (0.06)
XLNet	Expert-annotated	91.0 (4.70)	0.85 (0.06)
XLNet	Crowd annotated	74.2 (4.15)	0.54 (0.07)
Crowd ($\kappa = 0.007$)	-	73.9 (6.06)	0.61 (0.09)
Human Experts ($\kappa = 0.538$)	-	86.0 (4.40)	0.79 (0.07)

Note: Cross validation = 10 runs

3.3.4.1 Crowd-Trained Models Perform Poorly.

The transformer models trained with crowd-annotated data produced accuracies of 73.2% (3.85 s.d.) and 74.2% (4.15 s.d.) and F1 scores of 0.53 (0.06 s.d.) and 0.54 (0.07 s.d.) for BERT and XLNet, respectively (see Table 3.4). By contrast, we see a remarkable improvement in these results with the expert-trained BERT and XLNet models, which produced model accuracies of 89.1% (4.09 s.d.) and 91.0% (4.70 s.d.) and F1 scores of

0.82 (0.06 s.d.) and 0.85 (0.06 s.d.), respectively. We discovered that the enhancement in the F1 score is largely due to gains in the inter-rater reliability, which is the result of improvements in the quality of the training data between crowds and experts (see Fleiss' κ score increase from 0.007 to 0.538 in Table 3.4). We argue that inter-rater agreement is critical when working with annotated data from complex domains such as EV mobility. For reference, at the sub-topic level, values for Fleiss' κ range from -0.001 to 0.019 for the crowd, and 0.30 to 0.72 for the experts, which indicate considerable disagreement on the labeling task within a sample of 18+ adults representative of the U.S. population. See Experimental Procedures for details on human annotation experiments.

While sourcing strategies with online labor pools may be inexpensive, we find that the cost advantage does not justify the poor performance (F1 score 0.61, 0.09 s.d.). These results indicate that the use of low-cost crowd-sourcing approaches to build massive training sets are likely not feasible for large-scale implementation in this domain. This is in stark contrast to other deep learning domains, such as computer vision, where cheap, crowd-sourced training data can be easily acquired. For example, identifying sections of a road or public bus in an image is an easy task for the average person, but the average person cannot easily categorize the topics of EV user reviews. To provide an example of this, in our experiments, the review: *"...What an inconvenience when I need to drive to Glendale and I have a very low charge..."*, was cognitively difficult for general crowd annotators to correctly classify as *Range Anxiety*, even when annotators had unrestricted access to definitions and related examples. This was not the case for most experts. As a result, for these complex domains, expert-curated training data will be required for large-scale

implementations. In the next section, we compared the performance of our best classifiers using artificial intelligence versus human intelligence.

3.3.5 Possibility of Super-Human Classification

During hand validations of the transformers-based experiments, we noticed that some test data that were not correctly labeled by the human experts were being correctly labeled by the transformer models. This caught our attention as it indicated the possibility that BERT and XLNet could in some cases exceed the human experts in multi-label classification. In Table 3.5, we see that expert-trained transformer models performed about 3-5 percentage points higher in accuracy and 0.03-0.06 points higher in the F1 score as compared to our human experts. In Table 3.5 we provide 6 specific examples of this phenomenon where the expert-trained transformers do better than human experts. For example, exceeding human expert benchmarks could happen in multiple ways. It could be that the algorithm correctly detects a topic that the human experts did not detect (i.e. reviews 1 and 2 in Table 3.5); or that it does not detect a topic that has been incorrectly labeled by an expert (i.e. reviews 4-6 in Table 3.5); or that the sum of misclassification errors is smaller than that of human experts (i.e. reviews 3-6 in Table 3.5). We also provide quantitative measures in accuracy for these examples in Table 3.5.

Table 3.5 Examples where expert-trained transformers exceed human benchmarks (Ha et al., 2021)

				Expert-trained Transformers			
	Ground Truth	Human Expert		BERT		XLNet	
	Labels	Labels	Acc. (%)	Labels	Acc. (%)	Labels	Acc. (%)
1. *... unit says decommissioned but it will still release the charger after a long pause.*	Functionality	User Interaction	75	Functionality	100	Functionality	100
2. *Thanks very busy dealership but happy to allow use of qcdc*	Functionality, Availability, Dealership	Functionality, Dealership	87.5	Functionality, Availability, Dealership	100	Functionality, Availability, Dealership	100
3. *Charging on the quick charger - will be done by 12:15*	Functionality, User Interaction	Functionality, Location	75	User Interaction	87.5	User Interaction	87.5
4. *Went from 18-82% in 27 minutes! First time DC charging and met another nice Leaf owner who showed me how to use the machine. Thanks for the charge!*	Functionality, Service Time	Functionality, Availability, Location, User Interaction, Dealership	62.5	Service Time	87.5	Functionality, Service Time, Dealership	87.5
5. *The CHAdeMO charger does work. ... Nissan Hill had to move an ICE for me to gain access, but did so quickly. The CHAdeMO did not cost me any $ Charged quick! Don't hesitate to use.*	Functionality, Availability, Cost, Dealership	Functionality, Availability, Cost, User Interaction, Location, Service Time, Dealership	62.5	Functionality, Cost, Dealership	87.5	Functionality, Cost, Service Time, Dealership	75
6. *So the dealer had all of their cars being serviced parked in every spot including the quick charger. I called and asked them for at least access to the quick charger and they agreed but never did anything so I left and drove to Larry h nissan. I was willing to pay because I was in a hurry and obviously the Toyota dealer doesn't want my business.*	Availability, Cost, Dealership	Functionality, Availability, User Interaction, Location, Dealership	50	Availability, Dealership	87.5	Availability, Location, Dealership	75

Although a full investigation of superhuman performance for these transformer neural networks is outside the scope of the current study, we suggest this as an important future work. Evidence that artificial intelligence can outperform human benchmarks on multi-label classification tasks can benefit station managers and investors to be able to accurately

predict system problems or examine customer needs at high-resolution in ways not previously possible.

3.3.6 Applications for Local and Regional Policy

As EV consumer reviews data expands, we comment on the possibility to apply this computational approach widely to local and regional policy analysis. We note that previously, this type of extracted consumer intelligence has not been easily accessible to policy makers or governments due to the nature of unstructured data and issues with data access. For example, the U.S. Department of Energy's (DOE) Alternative Fuels Data Center maintains a list of all publicly accessible stations in the U.S. and Canada. This includes location information, such as station name, address, phone number, charging level (e.g. L1, L2 or L3), number of connectors, and operating hours with a developer-friendly API. However, these aggregated data sources do not typically include real-time usage or station availability, due to challenges with network interoperability.[16] This means that due to the presence of different charging standards by manufacturers and regional EV networks, there remain structural issues with sharing and receiving EV usage data between regions.

Recently, there has been a movement by a global consortium of public and private EV infrastructure leaders to promote open standards such as the Open Charge Point Protocol (OCPP; Open Charge Alliance 2020a) and the Open Smart Charging Protocol (OSCP; Open Charge Alliance 2020b). As these technology standards become more widely adopted, there will be a rapid increase in the amount of real-time data that can be shared with researchers and analysts. For instance, a growing number of digital platform providers have begun moving towards open data. These include platforms such as Open Charge Map,

Recharge and Google Maps. In the future, it should be possible to easily merge consumer reviews data with other spatial features and information. This could provide a wealth of commonly used features for analysis such as socio-economic indicators including population, income levels, educational attainment, age, poverty rates, unemployment, and affordability of nearby housing. Other important features could include transportation economic indicators, air pollution, health data, mobile phone tracking data, point of interest information, and local and regional incentives.

To provide an example of possible data insights for urban policy, we conducted a spatial analysis of metropolitan and micropolitan statistical areas (MSAs and μSAs). One of the dominant topics is *Availability*, which is predicted when a user reports whether a given charging station is available for use. In Figure 3.4, we visualize the spatial distribution of predicted station availability by U.S. census regions. To create this map, we merged the predicted review topics with counties based on shape files from the Office of Management and Budget's (OMB) 2013 specification of MSAs and μSAs. In the United States, there are 1,167 MSAs (population larger than 50,000) and 641 μSAs (population greater than 10,000), and 1,335 non-core-based statistical areas (population less than 10,000). To visualize model predictions, we standardized the predicted frequency of *Availability* topic into quantiles for each census region (West, Midwest, Northeast, and South), where 0-44%: *Rarely*, 45-69%: *Sometimes*, 70-90%: *A Moderate amount*, and over 90%: *A great deal* (see Figure 3.4). The map reveals areas with high and low predicted *Availability* consumer discussions in all core-based statistical areas.

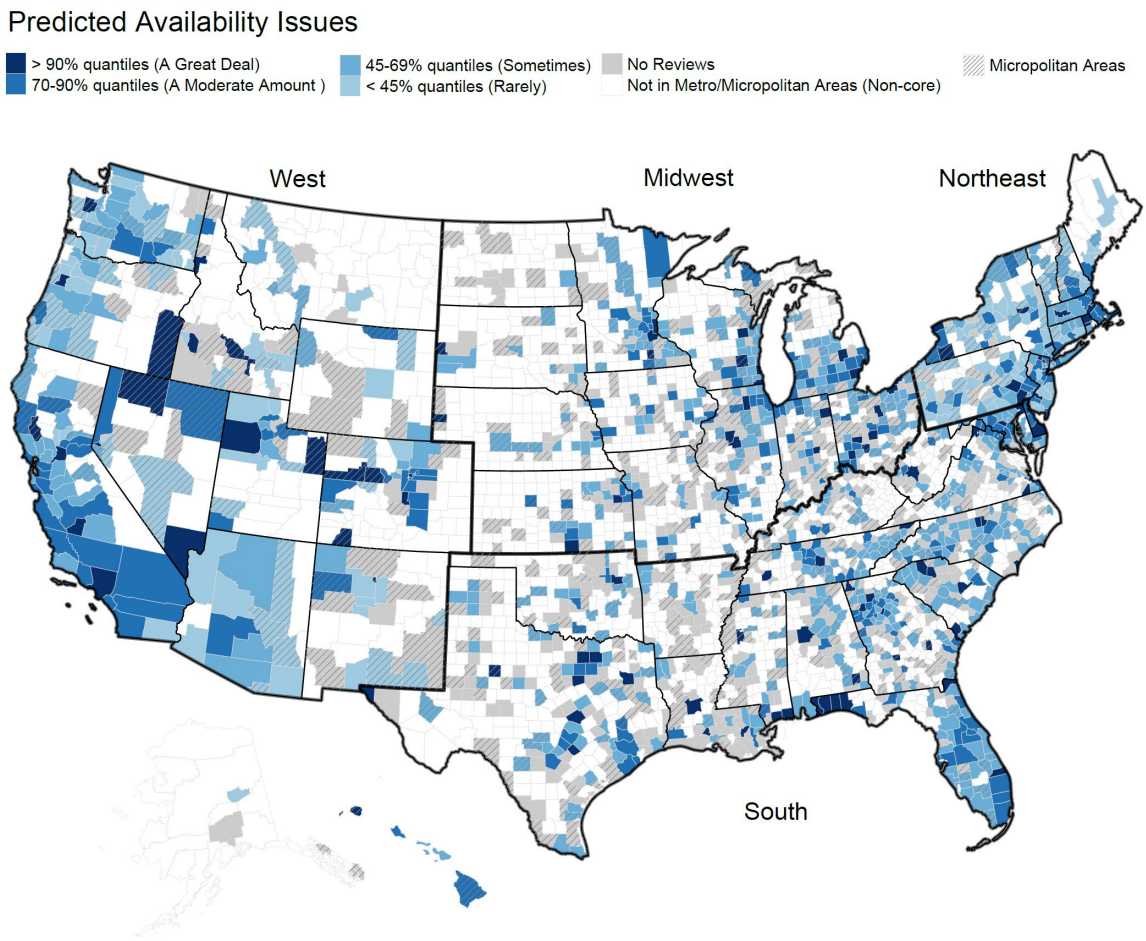


Figure 3.4 Predicted discussion frequency of station availability for US metropolitan and micropolitan statistical areas (Ha et al., 2021)

Using this approach, we find that predicted station availability issues are not necessarily concentrated in the large central metro counties (MSAs over 1 million population), but rather away from the city centers such as smaller μSAs of population less than 50,000. This is particularly true in the West (e.g. Oregon, Utah, Colorado, Wyoming, New Mexico) and Midwest (e.g. South Dakota and Nebraska) and Hawaii. By contrast, for the South (e.g. Texas, Alabama, Florida, North Carolina, South Carolina, Tennessee) and Northeast regions (e.g. New York, New Jersey, Massachusetts, Maryland, Pennsylvania),

we find the highest frequency of availability issues in the major MSAs for the period of analysis. One primary insight from this analysis is that μSAs could be under-served with regard to station availability. In additional analyses, we also used our methodology to detect whether a specific station is functioning. Based on the rate of consumers leaving reviews at charging stations across the U.S., we find that the deep learning algorithms can detect functioning of a certain station, daily. For further details of these estimates, see Supplemental Experimental Procedures. This type of detection could also be done with any of our introduced topics and with expanded sample datasets from network providers.

Given the proliferation of EV policies worldwide, this spatial analysis could be expanded globally. For example, in the European Union, policies such as Alternative Fuels Infrastructure Directives, or AFID (previously known as the Directive on Alternative Fuels Infrastructure, or DAFI; European Parliament and Council of the European Union 2014). In addition, the European Commission has supported implementation of fast charging infrastructure through the Trans-European Network for Transport (TEN-T) and Connecting Europe Facility Transport (CEF-T) programs (European Parliament and Council of the European Union 2014; TEN-T 2015). This type of national scale infrastructure expansion in the EU is part of an overall strategy by The European Union to reduce CO_2 emissions from the transportation sector by 60% by 2050 (European Commission, Directorate-General for Mobility and Transport 2011)

This capability to deploy accurate and more efficient deep learning models can be applied to evaluate other charging infrastructure roll-out policies that aim to increase the number of charge points, reduce charging congestion, promote vehicle-to-grid and overnight charging, as well as solar adoption (Kam et al., 2020). For recent reviews on how

charging behavior can guide charging infrastructure implementation policy, see Kam et al (2020). and McCollum et al (2018). Other applications that use artificial intelligence and NLP to discover hard-to-reveal patterns in unstructured data, especially those that merge spatial information, should generate fruitful areas of future inquiry.

3.4 Conclusion

In this study, we report state-of-the-art results for multi-label topic classification of consumer reviews in EV infrastructure. This represents a potential step change in our ability to aggregate data and insights for EV business model development and public policy advisory. Implementing automated topic modeling solutions has been challenging because of the technical nature of the corpus and training data imbalances. Our experimental protocols highlight the importance of the quality of training data annotations in the data processing pipeline. First, human expert annotators outperform the general crowd both in accuracy and F1 score metrics. This is due to improvements in the inter-rater reliability that is critical while working with data from complex domains. Second, improvements in training data quality also produce more accurate and reliable detection. This is seen in the approximate increase of 15 percentage points in accuracy and 50% improvement in the F1 score in the expert-trained transformer models as compared to the crowd-trained models (Table 3.4). Third, when the models are trained on top of high-quality expert curated training data, surprisingly the transformer neural networks can outperform even human experts. This indicates evidence of super-human classification on imbalanced corpora. As deep learning models have often been criticized for their black-box nature, we suggest

technical enhancements that focus on model interpretability as future work such as through the use of rationales (Zaidan et al., 2008), influence functions (Serrano and Smith 2019), or sequence tagging approaches (Nguyen et al., 2017) that can offer deeper insights on the models and the reasons for their predictions. This is an area of active research.

Further applications of methods that we propose particularly those that integrate artificial intelligence with real-time data and spatial analysis can greatly enhance new ways of thinking about infrastructure management as well as economic and policy analysis. Other opportunities abound.

3.5 Acknowledgments

We gratefully acknowledge funding support by the National Science Foundation (awards 1945332 and 1931980), a Microsoft Azure Sponsorship, and the Ivan Allen College Dean's SGR-C Award.

CHAPTER 4. INTERPRETING TRANSFORMERS ON GLOBAL EV CHARGING REVIEW CLASSIFICATION USING ATTENTION FLOWS

4.1 Introduction

In the previous chapter, we collected and studied 127,257 EV charging reviews from the US, that were generated during the period of 2011-2015. During this period, the global annual EV sales had increased from less than 50,000 in 2011 to 0.55 M in 2015 (ICCT 2-17). EV sales have consistently grown since then, reaching 0.8M, 1.2M, 2.1M, 2.2M in 2016 – 2019 periods each year (IEA 2020). Along with the increase in EV sales, the charging infrastructure has evolved substantially compared to the emerging years of 2011 – 2015. The number of public EV charging points in the US had increased from 3,410 in 2011, to 31,003 in 2015. Currently, it is estimated that there are 42,852 station locations in the US (DOE 2021).

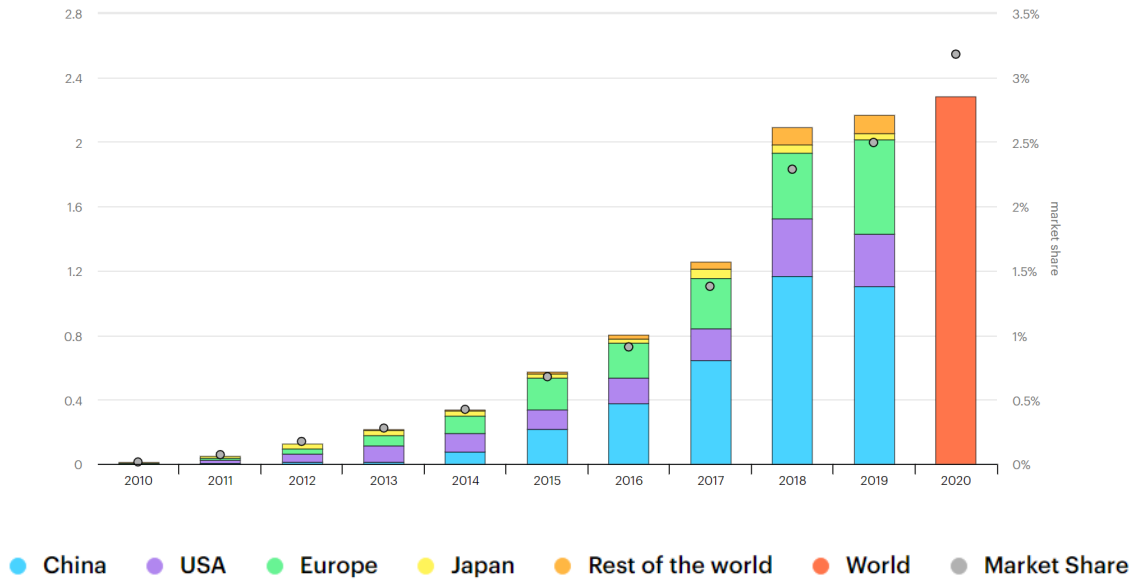


Figure 4.1 Global electric car sales by key markets, 2010 – 2020 (IEA 2020)

In the previous chapter, we have trained transformer models, BERT and XLNet as well as CNN and LSTM with training data annotated by experts, and saw promising results and showed how the models can benefit diagnosis of EV charging infrastructure performance related to the Availability issues. Given the steep increase in EV sales and charging infrastructure, more recent EV charging experience data are required to identify the current issues EV drivers may be facing.

The European Union has taken various actions to support electric mobility, implementing policies such as AFID. Further, implementation of fast charging infrastructure through programs such as TEN-T and CEF-T by the European Commission is part of an overall strategy by The European Union to reduce CO_2 emissions from the

transportation sector by 60% by 2050 (European Commission, Directorate-General for Mobility and Transport, 2011). As shown in Figure 4.1., European countries have become larger EV market than US since 2015, currently being about twice as large as the US as of 2019 (EIA, 2020).

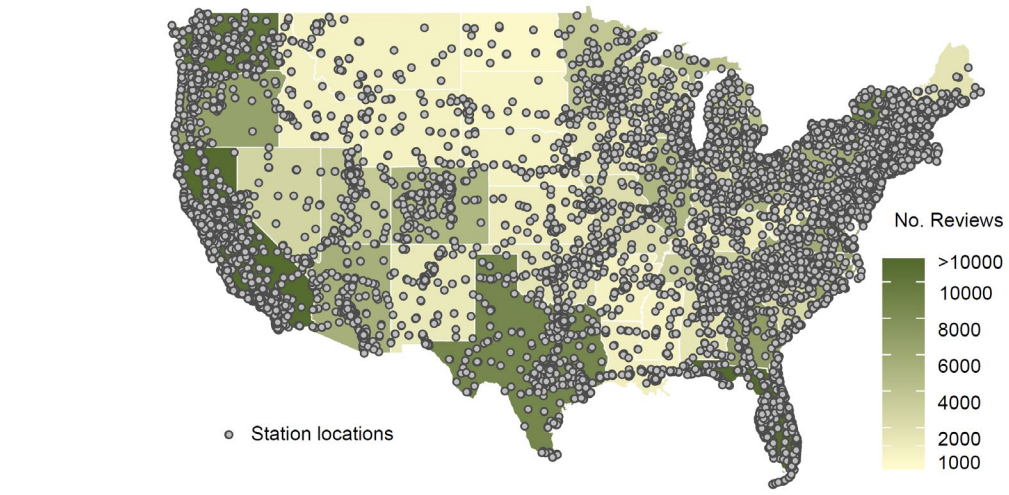
In this chapter, we collect the recent EV charging experience reviews generated from the US and EU in the 2016 – 2020 period, and evaluate the generalizability of the transformer models developed in the previous chapter, by implementing human expriments on data curation on the ground truth labels of both US and EU EV charging reviews. Further, using these transformer models, we provide interpretability on the behavior of the model's prediction on labels, using the attention flow. This chapter is part of a policy analysis research on the interoperability of EV charging network on the global scale including US and EU. This chapter serves as a computational set up for further large-scale research to evaluate performances of closed charging networks, such as the Tesla Superchargers, the world's largest closed EV network, comparing with the open networks which allow multiple network providers to access open protocols and allow different stations to communicate with each other.

The remainder of this chapter describes our data collection process with human experiments on the training data curation, presents a validation and generalizability of the transformer models on the larger dataset with expanded time period and the spatial range of the data generation. Then, we interpret the model's behavior on how it identifies the topics of each review using attention low. We then discuss how these computational results can be applied into an econometric analysis for policy analysis in the open networks versus closed charging networks in the US and EU areas.

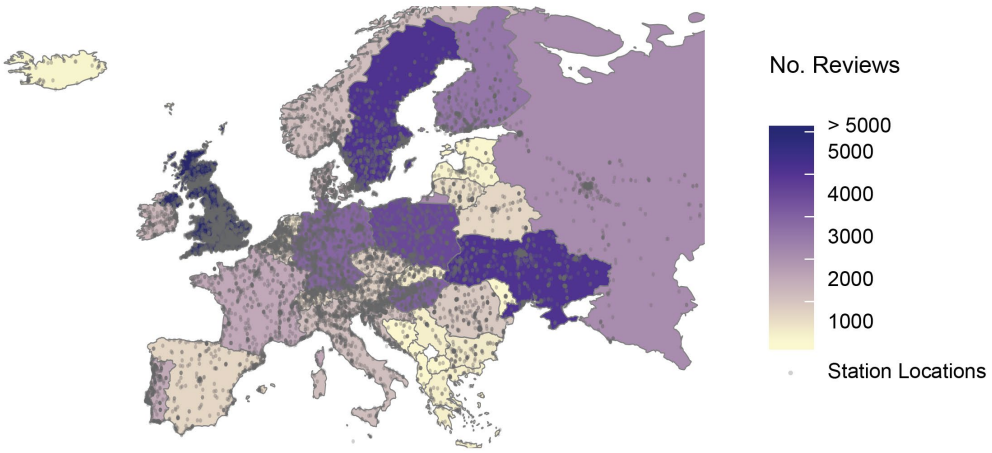
4.2 Methods

4.2.1 Electric Vehicle review data collection

Building on our existing data of EV reviews generated in the US during 2011 – 2015 period, we collected new data from US and EU, finding new stations. For US, we have collected total of 201,837 reviews during the period of 2014 – 2019, from total of 23,826 EV charging stations in US. As shown in Figure 4.2 (a), states in the west coast and the east coast had the highest numbers of reviews, such as California, Florida, Washington, and New York. From the European countries, we have collected total of 68,029 reviews from 17,194 station locations in 49 countries in Europe from the 2014 – 2020. These reviews were written in 53 different languages, with most popular languages of English, Russian, Polish, Hungarian, and Swedish. About 63% of the reviews were written in English, and the rest were translated into English language, using the Google cloud translation API. As Figure 4.2 (b) shows, reviews were collected largely from the United Kingdom (UK), having more than 27,000 reviews. The rest countries had less than 5,000 reviews, where Ukraine, Sweden, Poland, Hungary and Germany were placed at top countries in decreasing order. Although we have identified large numbers of reviews and station locations, note that these are not all of the stations that exist in both regions, but rather found by authors.



(a) United States



(b) Europe

Figure 4.2 Distribution of number of reviews and station locations

4.2.2 Human experiment: curating ground truth data

We conducted a human experiment to build a set of ground truth data for two main reasons. First, to validate the previously trained model presented in Chapter 3 and check

for generalizability throughout the different countries and the temporal differences. Second, retaining such high-quality data that can be used for experiments for interpretability evaluation examples and model performance augmentation.

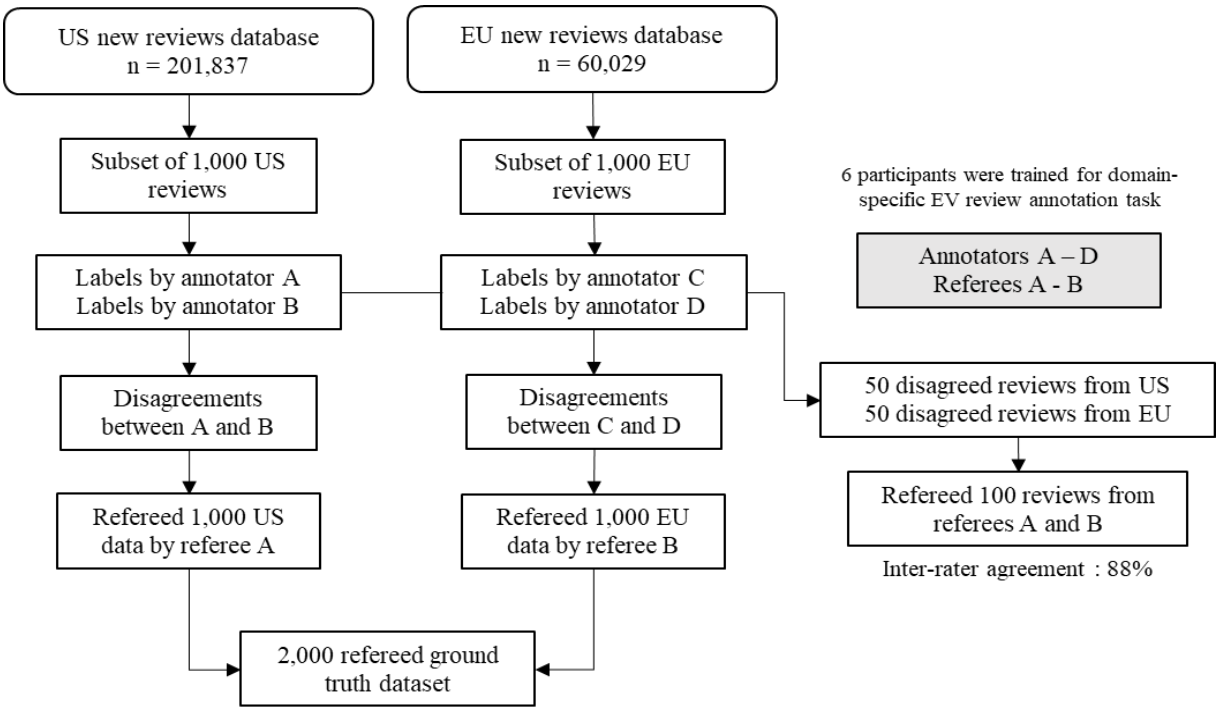


Figure 4.3 Ground truth data curation process

As shown in Figure 4.3, Curation of ground truth data curation is processed as follows: first, 6 participants were recruited and were trained by the authors for annotation of EV reviews based on the training manual developed in Chapter 3. Of the 6 participants, 4 of them are referred to expert annotators, and 2 of them are referred to as referees. Among the participants, referees were assigned based on the familiarity with the public EV charging behaviors. Secondly, 1,000 reviews were each randomly sampled from US and

EU review pools. This sampling process was done iteratively, making sure that sampled data had similar distribution in features including station quality rating, population of the station location in the county level, parking types, plug types, type of networks, and costs. Then, 4 expert annotators were divided into 2 groups of 2, where each group was assigned with either 1,000 US or EU sampled data for annotation. Table 4.1 shows the kappa values for main topics of US and EU data. Overall, the average agreement level is moderate agreement, while some main topics have very good agreement level (Cost and Charging Speed for both US and EU data). Range Anxiety annotation for US has the lowest level of agreement between 2 expert annotators. Given that Range Anxiety is a rare topic in the EV reviews, and the topic is usually identified with the context, without any specific keywords, it was expected, and needs further curation by referees. Also, agreement level on Dealership was among the lowest for both US and EU.

Table 4.1 Inter-rater agreement level measured by Cohen's kappa on 1,000 sample data of US and EU

	US	EU
Functionality	0.59	0.61
Range Anxiety	0.07	0.60
Availability	0.70	0.58
Cost	0.81	0.84
User Interaction	0.54	0.50
Location	0.63	0.66
Dealership	0.47	0.22
Charging Speed	0.84	0.81
Average	0.58	0.60

4.2.3 Transformer models

4.2.3.1 Bidirectional encoder representations from transformers (BERT)

BERT is a pre-trained contextual language model that leverage massive corpora such as the English Wikipedia and BooksCorpus to learn context from tokenized words (Devlin et al., 2019). BERT leverage neural network architectures with information feeding in a bidirectional context. The language models are fine-tuned on our domain specific multi-label classification problem using training data. For example, for a sample review, *"Fast charger working great!"*, BERT maximizes the conditional probability of the word context in the forward and backward direction as follows:

$$\mathcal{L}_{BERT} = logP(Fast \mid working\ great!) + logP(Charger \mid working\ great!)$$

Here, $\mathcal{L}_{BERT}$ refers to the log-likelihood functions for BERT. Figure 4.4 shows the representation of the BERT model architecture across the 8 topics of interest as a set of binary prediction outputs. For example, for the sample review shown *"Took 90 minutes . . . Ok. 5$ per charge"*, the model outputs 1 for Cost, Service Time topics, and 0 for the other topics. For seminal readings on BERT, see Devlin et al. (2019).

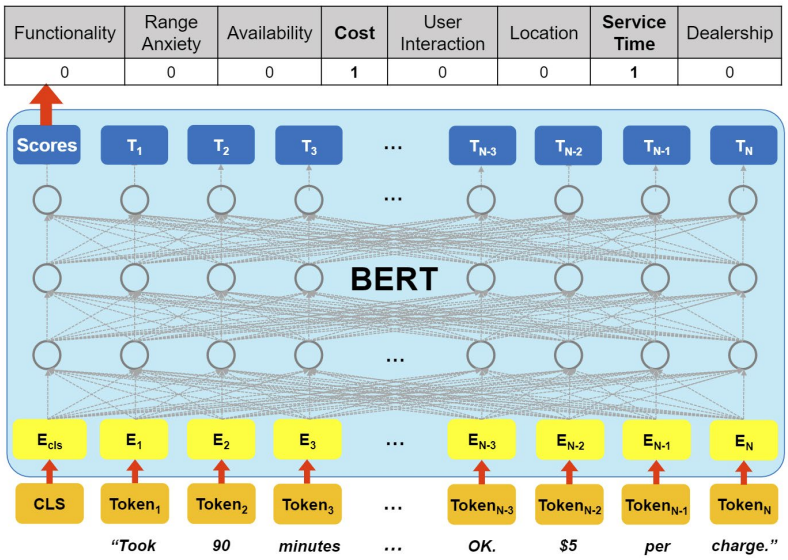


Figure 4.4 BERT model architecture (Ha et al., 2021)

4.2.3.2 Robustly optimized BERT-pretraining approach (RoBERTa)

Introduced at Facebook, Robustly optimized BERT approach, RoBERTa, is a retraining of BERT with improved training methodology, 1000% more data and compute power (Liu et al., 2019). While BERT uses masked sequence once in the pro-processing, RoBERTa is trained with duplicated training data by 10 times, so that each sequence was masked in 10 different patterns. Further, the model was trained for 40 epochs. With this process RoBERTa improves upon BERT's under-training issue that the same masking pattern is used for the same sequence in all the training process (Liu et al., 2019). Additionally, RoBERTa also adopted dynamic masking, where a masking pattern is generated every time a sequence is fed to the model, and slight improvement was achieved.

While BERT provides a robust baseline, RoBERTa has shown improvements on the performance metrics. In this study, we use both transformers for the multi-label classification problem for the EV reviews to serve as basis for interpretation.

4.2.4 Attention flow for transformer interpretation

The attention flow is a method for approximating the attention to input tokens by using the attention weights as the relative relevance of the input tokens. It was proposed by Abnar and Zuidema (2020), considering the attention graph as a flow network. Using a maximum flow algorithm, it computes maximum flow values, from hidden embeddings to input tokens (Abnar and Zuidema, 2020). By quantifying attention flow, a new set of attention weights can take token identity problem into consideration and can serve as a better diagnostic tool for visualization and debugging (Abnar and Zuidema, 2020).

Visualization of attentions in transformer models is the easiest and most popular approach to interpret a model's decisions and to gain insights about its internals (Vaswani et al., 2017; Dehghani et al., 2019; Chen and Ji, 2019; Coenen et al., 2019; Clark et al., 2019). While it offers plausible and meaningful interpretations, attention does not equate with explanation (Abnar and Zuidema, 2020). Quantifying attention flow can give better explanations on the models' internal behaviors, with simple assumptions by approximating information flow in a model with the attention weights.

In this chapter, we use attention flow method to evaluate the inner behavior of the transformer models on our domain-specific text data and its complex multi-label classification task. By demonstrating the interpretability of the transformer models, we address the black box problem of deep learning methods such as the transformer models

for better management of the risks and harms that arise from lack of transparency of such models.

4.3 Results & Discussion

4.3.1 Transformer models

As shown in Table 4.2, we find similar performances between BERT and RoBERa. With the overall performance measured by macro-average F1 score, RoBERTa slightly outperformed BERT, having 0.72 and 0.71 as F1 scores respectively. In the topic level, while RoBERTa showed higher F1 scores in some topics such as Cost, Location, Charging Speed, and Dealership, BERT has outperformed RoBERTa in the rest of the topics. For the Range Anxiety topic, which seem to be the hardest topic to correctly predict by RoBERTa, BERT had substantially higher F1 score from BERT model, by about 0.23 points. Given the higher standard deviation of topic level F1 scores from RoBERTa and the marginal outperformance, we find that BERT is a more computationally cost-effective model for our dataset.

Table 4.2 Multi-label classification performance of transformer models

	BERT			RoBERTa		
	Precision	Recall	F1	Precision	Recall	F1
Macro-averaged	0.72	0.70	0.71	0.77	0.69	0.72
Functionality	0.83	0.85	0.84	0.88	0.82	0.85
Range Anxiety	0.67	0.59	0.63	0.53	0.40	0.46
Availability	0.79	0.75	0.77	0.85	0.76	0.79
Cost	0.83	0.89	0.86	0.86	0.90	0.88
User Interaction	0.6	0.42	0.50	0.69	0.45	0.54
Location	0.75	0.78	0.76	0.83	0.82	0.82
Charging Speed	0.72	0.76	0.74	0.83	0.82	0.82
Dealership	0.56	0.55	0.56	0.69	0.61	0.65

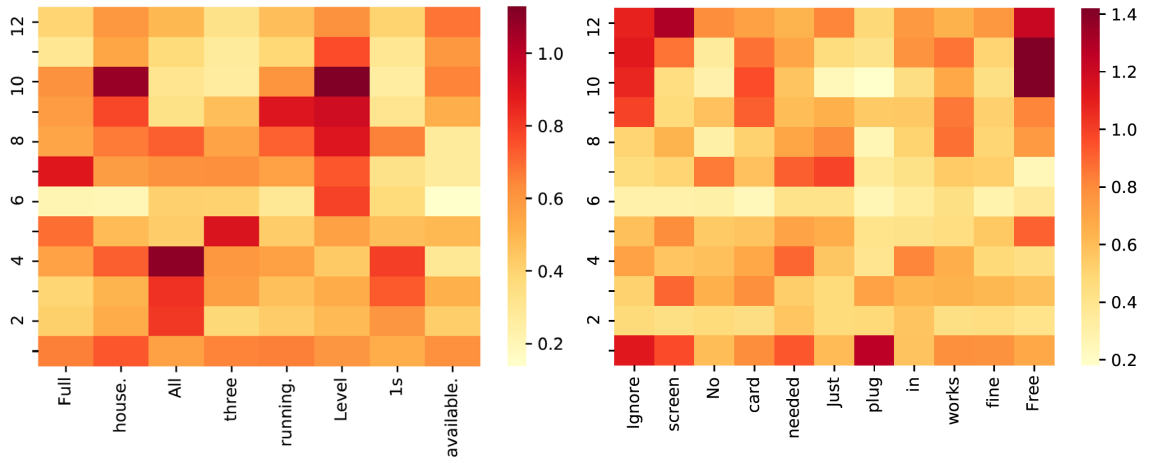
Further, when training the same BERT model with only US data collected from 2011 - 2015, the macro-F1 score was 0.73, which is only 0.02 points higher than the combined training dataset. Therefore, we also claim that the model is well generalized with international and translated review data, with tradeoff of 0.02 points in F1 score.

4.3.2 Interpretation using attention flow

Based on the fine-tuned BERT model, heatmaps of the quantified attention flow are showed with interpretation and discussion. First, we discuss attention flow weights of Functionality, Availability, Cost, Charging Speed and Location, as the model had F1 scores of close to or higher than 0.75 for these topics. Then, we discuss how the model behaves on predicting other relatively lower performing topics.

4.3.2.1 Functionality

Figure 4.5 shows the attention flow heatmap of review examples predicted as Functionality, as well as other topics such as Availability and Cost. In Figure 4.5 (a), we can wee that the attention flow weights are darkest for words such as “full”, “house”, “all” and “level”. Towards the last layer, we see that “house”, “level”, and “available” have highest attention flow weights. Here, “level” word is the predictor for Functionality topic, as charger types are often referred to as level 1/2/3. With the example shown in Figure 4.5 (b), we can easily see that the first two tokens have highest attention flow weights, as well as the last token “free”, leading to prediction of Functionality and Cost, respectively.



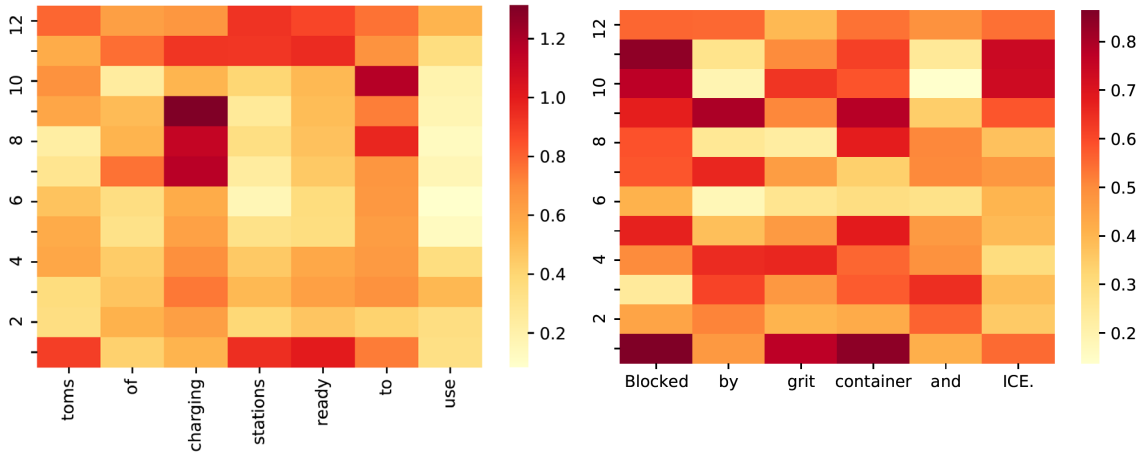
(a) Functionality and Availability (b) Functionality and Cost

Figure 4.5 Attention flow heatmap of review examples predicted as Functionality

4.3.2.2 Availability

In Availability topic examples shown in Figure 4.6 (a), it was interesting to find that what we believe to be a typo of “tons”, which was instead written as “toms”, eventually

was one of the highest attention flow weight in the last layer, along with “stations ready” phrase. However, the model also falsely predicted Functionality, and it may be due to “charging” token having the highest flow weight. On the other hand, example shown in Figure 4.6 (b) is rather straightforward to interpret, where some obvious words that indicate availability issues are shown to have the highest weights, for example, “blocked” and “ICE” in the later layers.



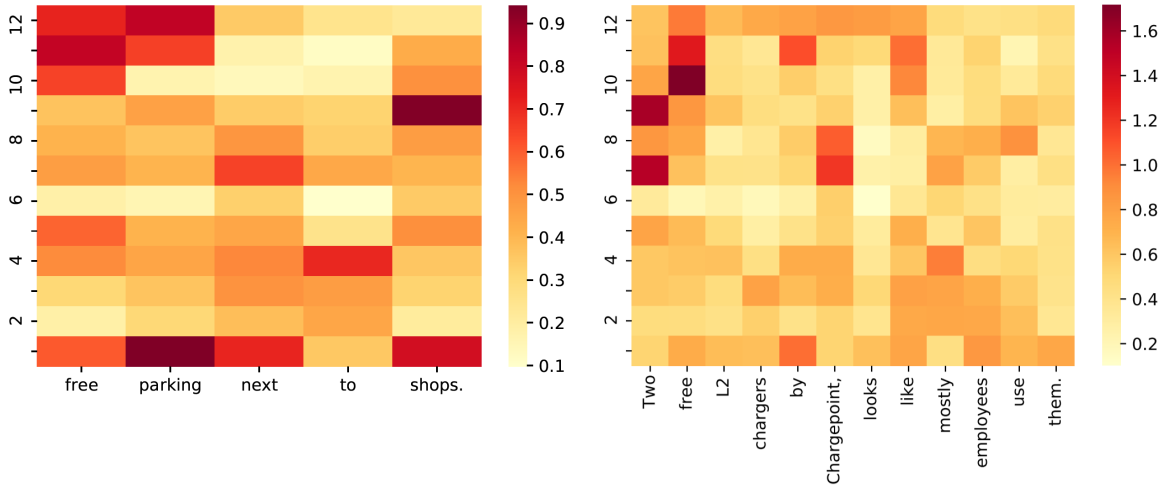
(a) Availability and Functionality (b) Availability

Figure 4.6 Attention flow heatmap of review examples predicted as Availability

4.3.2.3 Cost

Cost topic is often predicted with other topics such as Location and Availability. Example shown in Figure 4.7 (a) implies that “free parking” is the strongest indicator for Cost topic, and “shops” for Location. For the other example in Figure 4.7 (b), the first two tokens, “Two free”, had the highest weight values which is straightforward. For this

example, Availability is also predicted, implying that “two” token also is an indicator for Availability, discussing how many chargers are available at a station location.

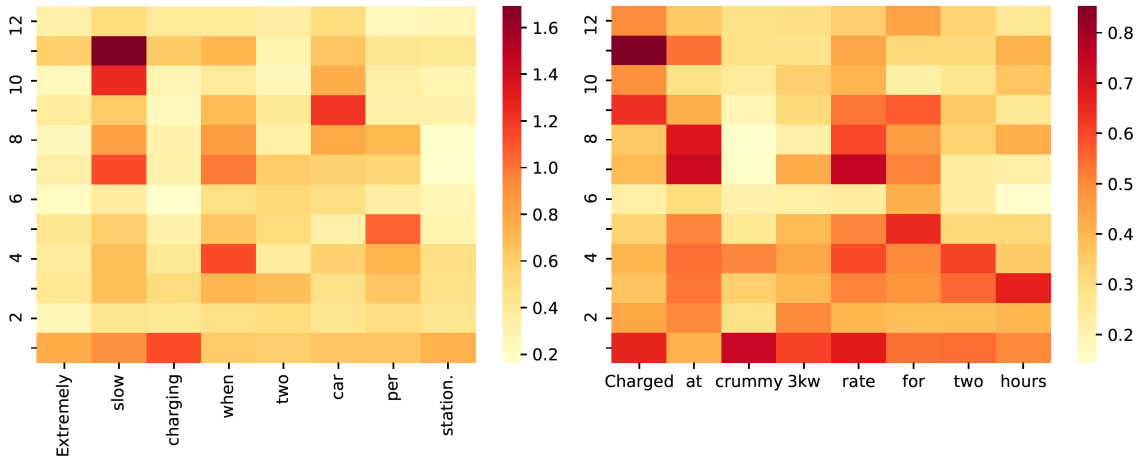


(a) Cost and Location (b) Cost and Availability

Figure 4.7 Attention flow heatmap of review examples predicted as Cost

4.3.2.4 Charging Speed

Charging Speed topic is probably the most straightforward topic the model was able to predict easily. As shown in Figure 4.8 (a), “slow” token has the highest attention flow weight, correctly indicating the prediction towards this topic. In Figure 4.8 (b), tokens such as “charged at”, “rate” have the highest weights.

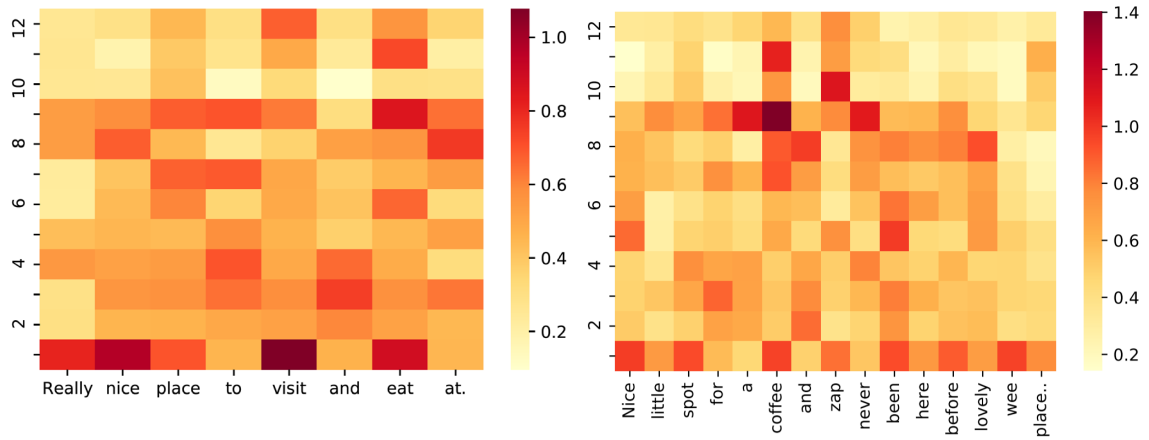


(a) (b)

Figure 4.8 Attention flow heatmap of review examples predicted as Charging Speed

4.3.2.5 Location

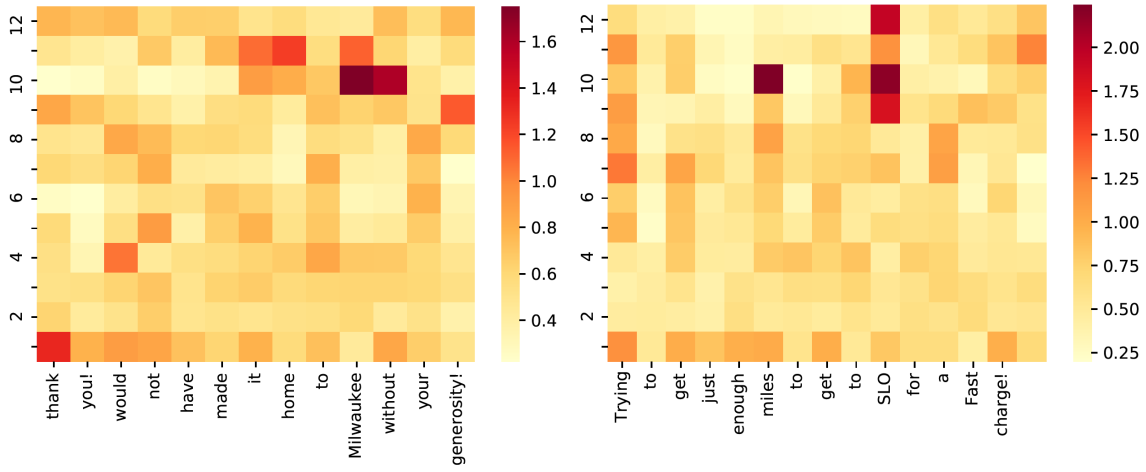
Although Location topic had F1 score of 0.76, tokens that contributes to Location prediction are rather diverse, instead of having certain keywords. For example, “visit” and “eat” tokens are the highest indicators for prediction as shown in Figure 4.9 (a). On Figure 4.9 (b), we see that “coffee” and “zap” tokens are the contributing tokens. From these two examples, we find that for Location topic, the model focuses on the actions one can take in the charging station location, and what one can get from here. This is also seen from example Figure 4.7 (a), where “shop” was the indicator token for Location.



(a) (b)

Figure 4.9 Attention flow heatmap of review examples predicted as Location

As the model performance on Range Anxiety, User Interaction, and Dealership were not as good, the heatmaps findings were not as clear as the discussed topics. For example, a very important topic among the EV policy discussions, Range Anxiety examples were not as consistently showing indicator tokens from the model. As shown in Figure 4.10, we see that the tokens that represent destinations from a trip, are contributors of Range Anxiety prediction. These tokens are "home", "Milwaukee" from Figure 4.10 (a), and "SLO" from Figure 4.10 (b). This is rather unexpected behaviour of the model compared with human decision, where "made it home", or "enough miles" were thought to be the largest contributions for Range Anxiety prediction. The model does highlight "home", "miles", but these do not exactly match human's perception. Given the relatively lower performance on these three topics, we leave it as future study to improve the model performance on these topics by collecting more quality training data and evaluate the internal model behaviors on these topics.



(a) (b)

Figure 4.10 Attention flow heatmap of review examples predicted as Range Anxiety

4.4 Conclusion

In this chapter, we collected more recent EV charging experience data for analysis of charging infrastructure behaviors in a global scale, including the US and the European countries. Then, we validated the transformer models developed in the previous chapter that the trained BERT model generalized for the global scale data, by testing on ground truth data from the newly collected data. These new labeled data are curated by human experiment, with training the annotators and going through the referee auditing process. Then, we further provide interpretations on the BERT model behavior on each of the main topics of EV reviews. We find that majority of the topic predictions from the model provide clear explanation on how the model identifies each topics, overcoming to some extent the black box criticism that neural networks often receive. Artificial intelligence when applied

to policy domain, providing clear interpretation and transparency of such deep learning models are very important, to minimize ethical harms it may cause.

4.4.1 Future study

This chapter serves as a computational model evaluation process as part of an ongoing global and large-scale research to evaluate performances of closed charging networks, such as the Tesla Superchargers, the world's largest closed EV network, comparing with the open networks which allow multiple network providers to access open protocols and allow different stations to communicate with each other.

In the computational perspective, striving for performance improvement of the models for the minor topics – Range Anxiety, User Interaction, and Dealership--, is a next step for policy analysis in various aspects of EV charging infrastructure. To achieve improvement, further high-quality training data should be collected. This will lead to clear explanable transformer model to classify EV reviews. Also, with the increasing number of labeled data with sub-topics, we expect to investigate deeper into each topic.

For policy analysis, econometric analysis such as the fractional response model, is in process of development. Using this model, we can evaluate the impacts of network technology policies on the sentiments of the EV charging experience, adjusting for observable station and location characteristics.

CHAPTER 5. CONCLUSION

Across my doctoral research I led efforts at the intersection of computational data analytics, infrastructure management, and energy policy analysis to demonstrate innovative approaches of deep learning and NLP for large-scale data analysis in this domain.

Collectively, these three studies contribute to the body of knowledge as follows. First, I demonstrate the design of experiments to curate high-quality training data. Second, I use state-of-the-art models and tailor them to the new domain, showing high performance and finding evidence in computational learning of domain specific terms that can't be learned from ordinary corpus. Third, overall, I demonstrate the framework of applying supervised deep learning for sustainable policy analysis.

The most important contribution from my dissertation, is the data curation part, where I demonstrate the process of obtaining high quality data, and showing deep learning models can perform well on these highly targeted domain specific tasks. I show it for two relatively different cases, where rigorously coded data with qualitative study standards can perform well with only small size, and where carefully curated design for a larger scale training data collection can train state-of-the-art transformer models. The results of this dissertation research represent a potential step change in our ability to aggregate data and insights for large-scale public comments data to support efficient policymaking process.

APPENDIX A. CODEBOOK FOR THE CLEAN POWER PLAN COMMENT DATA ANNOTATION

A.1 Support/Other/Oppose labels for Clean Power Plan comment data

There are 3 categories, with no subcategories. The 3 categories are mutually exclusive.

A.1.1 Support: 1

- Supporting implementation of the Clean Power Plan
- Opposing to repeal of the CPP
- Not Support: the comment mentions difficulty of breathing, protecting the planet, saving environment for future generations but has no explicit mention of support for CPP
- *"The last time I checked, we didn't find another habitable planet, and I guarantee you that even when Elon Musk built his ship, he is not going to take us on it to another planet. When many of us live in comfortable homes, have immediate access to clean water and have privilege to breathe fresh air, it's difficult to comprehend how other people are suffering and that we are in the disaster state."*
    i. The speaker is probably pro- Clean Power Plan, but there is no explicit mention of it in the text.

A.1.2 Oppose: -1

- Opposing the CPP

- Support repeal of CPP
- Not Oppose: comment gives signals anti Clean Power Plan by mentioning disadvantages, faultiness of CPP implementation but no explicit mention of oppose to CPP

A.1.3 Other: 0

- No explicit mention of support/oppose of CPP
- Main topic of comment is about something else, with no explicit mention of support/oppose to CPP
- Main topic of comment is too diverse, without in-depth opinion/knowledge that leads a specific conclusion of whether this person is supporting/opposing to CPP.
- Can assume the speaker is pro or con for climate change protection or coal industry, but no explicit mention in the text
- *"my contention is here today that everybody i know in this room and a lot more people out there, we all hear the water. it's not easy to hear because it's not easy to really know in your heart. in the work that i do, i've met so many people. i just had lunch with somebody two weeks ago that after he saw my presentation on climate change sat at the table and just sobbed. i hope it's not too late. i appreciate what you guys are doing, and it's a first step. we need to address the methane. this morning somebody said, and this worries me, that, you know, by focusing on the goal, the methane will just go. and it is worse for the climate than any coal. and the other thing that i hope we start hearing about is a carbon tax because that's where it's going to change everything is having a carbon for your dividend. so just in closing i*

just want to say that, again, as a mother, i do believe, and this comes from our children's trust, it's an old environmental law."

   i. Does not give clear signal on any mentioned topics – mentions future generation, methane, carbon tax

- Main topic is about climate, but it is irrelevant to CPP, and doesn't mention speaker's stance on the CPP.
- *"i want to talk today about a couple of issues that i have been researching since then. the first one has to do with the pattern we see when it comes to sea level rise. when you look at the time in which i have done most of my thinking about this issue, you see that beginning in about we began to see this terrible change in the seasonal pattern. the ice melt in the artic –- in the sea ice. the sea ice doesn't have anything to do with sea level rise, but the energy that goes up there is pretty substantial"……. "so the second problem that i have has to do with longevity, and i have been doing a lot of work on that lately. the problem with longevity is that when we make our estimates, we use a way of looking at the data that ignores the fact that as you get younger cohorts, they have overall a better life expectancy in the long run……."*

   i. Mentions climate impacts, but no argument basis that is related to CPP

A.2 Topic labels for Clean Power Plan comment data machine learning model

There are 4 main topic categories and 13 subtopic categories in total.

Multiple main topics and subtopics can apply to a comment. It is possible that only main topic is selected, without selected subtopics.

Main Topics	Subtopics
Environmental Impacts	Climate Climate Pollution Health Non-climate Pollution Extreme Events
Economy	Jobs Costs Future Economy
Resources	Coal Natural Gas Clean Energy
Ethics	Future Generation Justice Stewardship

A.2.1 Environmental Impacts

Comments that mention general environmental impacts apply to this main topic. It can be about climate, health, pollution, and extreme events. Multiple subtopics can apply to one comment.

A.2.1.1 Climate

- Mention of the following words: climate change, sea level rise, temperature, global warming/temperature
    i. *"Clean Power Plan might have limited rising water if fully implemented"*
- Mention of climate being put at risk, crisis, threat
    i. *"Repealing the Plan means ignoring the reality of the climate crisis"*

- Mention of emissions that are not directly harmful to human body, but causes climate impacts: green house gas (GHG) emissions, carbon dioxide (CO2) emissions, including methane
- Mention of emission with no specification
    i. *"lawmakers are finally beginning to address this carbon footprint"*
    ii. *"helping to avoid millions of additional tons of carbon emissions"*
- Not Climate: mention of related words in description of different context or referencing a third party (not in the environmental impact context)
    i. *"I have a master's degree in climate science"*
    ii. *"i mean, when XX percent of the scientists are telling us that climate action is actually happening because we are contributing to it,"*
- Not Climate: mention of rule or goal names:
    i. *carbon intensity standards*
    ii. *determining how it will hit its carbon pollution target*
    iii. *"i'm here to urge you to please work towards a larger national emission reduction goal"*

A.2.1.2 Health

- Mention of diseases, such as asthma, lung cancer, health, death
- Mention of health of ecosystem
    i. *"the forests that surround helena are now full of dead and dying trees"*
    ii. *"Healthy environment"*
- Mention of health in the context of community

i. *"threaten the wellbeing and safety of my community"*

- Not Health: mention of illness or sickness with no related context or referencing a third party
    i. "we in west virginia hear the words of president obama's administration saying, "coal makes us sick. coal is our worst nightmare"

A.2.1.3 Non-climate Pollution

- Mention of pollution, pollutants that are directly harmful to organisms, human
- Mentions that imply pollution
    i. "we have done a very good job of controlling all gaseous by products that comes with burning coal and other Fossil Fuels to the point that this region is emitting lower emissions than any required standard"
- Mention of sulphur dioxide, soot, coal ash, smog, change in quality of air/water/land
    i. "In using modern technology, it can do so with outstanding air quality."
- Not Non-climate Pollution: pollution that refer to green house gas emissions
    i. Methane (green house gas, thus climate)
    ii. *As the EPA is aware, humans emitting greenhouse gasses into the air is the primary cause of climate change. We've witnessed the effects of this pollution in California in the form of record heat, record droughts* – **falls under climate, not non-climate pollution tag**
    iii. The Clean Power Plan is our best chance to limit carbon pollution from power plants, which is the largest source of such pollution

iv. *I am speaking about frac sand mining today because the epa's proposed carbon pollution standards address only a small segment*

- Not Non-climate Pollution : mention of pollution with no related context towards harmful pollutants

i. *Children are not little adults. They are more sensitive to environmental changes because of their physiology, and more likely to be exposed to environmental contaminants because of their stage in development.*

A.2.1.4 Extreme Events

- Mention of climate related or general extreme events – extreme heat, heatwave, temperature, flood, fire, hurricane, hazards
  i. *"protect our citizenry from avoidable environmental hazards, including catastrophic climate disruption"*
  ii. *"she mentioned that in the last few years the storms that have hit the coastline come more frequently and with greater force."*
  iii. *"they've seen extreme rain events, unseasonably hot or cold weather, even for erie, and drought."*
- Not Extreme Events: extreme events that are not-climate related, i.e., fire breakouts by pollution
  i. *Cuyahoga River fire – caught on fire due to industrial pollution* – Should be labeled as pollution, instead of extreme events

A.2.2 Economy

Comments that mention domestic or international economy apply to this main topic. Jobs, costs, future economy mentions apply to this topic.

A.2.2.1 Jobs

- Mention of jobs, creation of jobs, losses of jobs, employees
    i. *"Miners are being laid off at no fault of their own"*
    ii. *"KCP&L approximately 2700 employees including about 1,600 represented by three local unions of the International Brotherhood of Electrical Workers (IBEW). The employees live and work in the Kansas City area"*
    iii. *"this failure comes at an extreme cost for those working in coal and coal related industries"*

A.2.2.2 Costs

- Mention of price of energy, implementation cost of the Clean Power Plan, cost for transitioning to renewable energy
    i. *"that will mean higher energy prices for those who can least afford it"*
    ii. *"if you make us do this, we're going to raise our rates"*
    iii. *"not only are prices for both solar and wind power dropping rapidly"*
    iv. *"the price of solar pv has dropped more than percent since and was percent less at the end of to xx."*
    v. *"producers of carbon intensive grids will be burdened"*
- General mention of cost-benefit analysis
- Mention of costs as consequence of CPP implementation – (i.e., health care costs, business impacts)

i. *increases our overall health care costs*

ii. *"efforts to remove this costly and biassed rule"*

iii. *"in addition to a negative economic impact, climate change is also threatening"*

- Mention of costs related to climate change and events

i. *"it caused over $ billion in damage to homes, businesses, and infrastructure"*

ii. *"Consider for example Malibu mudslides, the Napa wildfires, Southeast Asian billionaires and tycoons who have lost everything in monsoons, Hurricane Sandy, Hurricane Harvey, millionaire losses"*

- Not Cost: mention of cost not in context of economic cost

i. *"the cost of failing to adequately deal with climate disruption caused mainly by carbon pollution are immense."*

A.2.2.3 Future Economy

- Mention of economic growth/decline from impact of CPP implementation/repeal

i. *"we can replace the retired dirty coal plants with geothermal, solar and wind energy, and the growth in these industries will more than replace the jobs lost from the coal industry"*

ii. *"The carbon the Clean Power Plan has become an engine of the economy in the U.S. and the world in the coming decades."*

iii. *"they are not going to transition to a new and better economy, energy economy"*

- Mention of growth/decline in an industry, economy, GDP, economic competitiveness, sometimes indicated by growth/decline of future jobs

i. *"don't believe that epa's cost analysis has fully incorporated some of these kinds of agricultural impacts and i urge you to do a more careful job of fully analyzing all of the costs of climate change and the cost of inaction and doing nothing."*

ii. *"job-killing regulation"*

iii. *"create thousands of jobs"*

iv. *"more competitive in the clean energy sector"*

v. *"building vibrant rural communities"*

vi. *"Coal is not going to return; the markets have moved on"*

vii. *"utility scale wind, solar, unsubsidized, are already cheaper than naturla gas"*

viii. *"And the fact is that there are many more jobs in renewable energy already. And there can be many, many more"*

ix. *"jobs in the future"*

x. *"fossil fuel workers who are going to lose their jobs because not because of regulations in the Clean Power Plan, but because of the very free market that Trump claims to worship and claims to be in favor of is the one that's eliminating them."*

xi. *"And for me, it will help ensure that my family can continue their successful farming business"*

xii. *"at the national level a revenue neutral carbon tax would have a correspondingly greater positive impact and this includes . XX million jobs added, $XX trillion added to the gdp, lives saved while cutting carbon emissions by XXX percent."*

xiii. *"as if they are not going to transition to a new and better economy, energy economy, because they will*

xiv. *"willing to sacrifice what's going to amount to trillion of our aggregate GDP over the next years"*

xv. *"negatively impact the trona and soda ash industry's ability to compete globally,"*

xvi. *"pennsylvania has an opportunity to develop a strong, all inclusive plan that moves us forward, to once again be the leader of energy throughout the world."*

- Mention of new industry, best practice for carbon utilization/sequestration, test facilities, growth/emergence of new technology, industry

  i. *"hydropower can double its contribution towards a clean energy economy, helping to avoid millions of additional tons of carbon emissions, serve as a cost effective compliance option, create hundreds of thousands of new jobs, strengthen our national infrastructure, and increase the reliability and resiliency of our electric grid, all priorities for the current administration."*

  ii. *"There's lots of examples of utilizing carbon from burning coal."*

  iii. *"the implementation of renewable energy in these states should be seen as a gateway to new industry, not a blockade to financial success."*

  iv. *"Sheldon Station near Hallam would be the first utility-scale hydrogen powered generator in the U.S., and is expected to produce 125 megawatts of clean electricity."*

- Not Future Economy: mention of economy in the past – main topic Economy should be selected
    i. *"the U.S. direct economic hits from 19xx to today, average and direct economic hits was xx billion"*
- Not Future Economy: mere mention of economy without specific drive of growth/decline of economy
    i. *"Agency's statutory authority will provide regulatory certainty while promoting the environment and the economy"*

A.2.3 Resources

General resource: *fossil fuel* (only to main topic and does not apply to any subtopics.) – such as oil

A.2.3.1 Coal

- Mention of coal, coal mine
- Mention of specific names of coal power plant
    i. *"Plant Scherer"*
- Not Coal: mention of coal in context of pronoun: for example, name of an institute
    i. *"I serve as president for the rocky mountain coal mining institute"*
- Not Coal: mention of coal in referring a third party records:
    i. *"we in west virginia hear the words of president obama's administration saying, "coal makes us sick. coal is our worst nightmare""*

A.2.3.2 Natural Gas

- Mention and implications of natural gas, methane, gas pipes

i. *"shifting to fracking continues to emit carbon from venting and flaring"*

A.2.3.3 Clean Energy

- Mention of solar, wind, hydropower energy, renewable energy, clean energy, ethanol
  i. *"The jobs in the clean energy sector are going up significantly"*
  ii. *"i would like to request that the epa encourage all states heavily reliant on coal, such as Georgia, to increase the proportion of energy from renewable sources, such as solar and wind"*

Not Resources : General mention of power plants with no specific resource for it.

A.2.4 Ethics

A.2.4.1 Future Generation

- Mention of children, grandchildren, future generation, young people
  i. *"there's a family with two young children"*
  ii. *"I worry that my children will be less healthy and could die prematurely."*
  iii. *"Most of us adults will not live to see the gravest consequences of not acting on climate change, but these young people will."*
  iv. *"My son, Shanti, and the children of our Congressman Lacy Clay are among many who suffered from asthma while growing up in the St. Louis area."*
  v. *"i'm a grandmother and i'm going to speak as a grandmother who would like for my grandchildren and their children to inherit a safe planet"*
  vi. *"Allowing coal burning power plants to pollute the air, sicken children, and dump carbon into the atmosphere, disrupting the climate is not protecting human health and the environment."*

vii. *"I am saying that it has impacted where our children are going"*

- Mention of status as parent or grandparent
  i. *"I am here as a mother of 4 children"*
- Not Future Generation: No explicit mention of children, future generation
  i. *"the terrible consequences that this proposal would have for my family"*

A.2.4.2 Justice

- Mention of justice, environmental justice, bias, equity
  i. *"with all revenues returned on an equitable basis to households"*
  ii. *"It was biassed from the beginning"*
  iii. *"We need a just transition to renewable energy that doesn't abandon our miners and affect the communities."*
  iv. *"We also do grave injustice to the members of the coal industry"*
- Mention of unfairness/fairness/threat/support/privilege to certain demographics community – race, age, income, location, jobs
  i. "*do we need to make sure that our workers aren't left behind and that the clean energy transition works for everything?"*
  ii. *"haves and have-nots"*
  iii. *""they are a small group of the people who are more vulnerable to the impacts of changing climate in Colorado"*
  iv. *"but we can clean up the atmosphere and improve the quality of life for all Georgians"*
  v. *"So even the most privileged, even the millionaires, even the billionaires amongst us, we may seem to live in a different world"*

vi. *"Farmers don't have the luxury of sitting indoors near air filters."*

vii. *"support for the working class communities and communities of color across the country that are hit first and worse by the impacts of energy pollution and climate change"*

viii. *"the price of food will likely increase, and the poor are almost always the hardest hit when we have harsh weather related disasters"*

ix. *"all of which hurts Wyoming's middle class families and workers"*

x. *"documents that are crafted to cloud the true intent from average citizens who don't have the time nor expertise to verify the true intent"*

xi. *"do it for the poor, the sick, the disenfranchised"*

xii. *"that does not mean that we can neglect mineworkers, their families and communities like my hometown. we need to commit to helping them transition just as we have helped tobacco farming and logging communities transition in the past"*

xiii. *"they fear the loss of jobs in a way of life, and at the same time we look at this rule, we must be concerned for coal miners and power plant workers and find a way to address their fears"*

xiv. *"ignorant about the real world of energy development and environmental protection or deeply hypocritical in its commitment to save coal dependent communities like Gillette"*

- Mention of unfairness/unlawfulness about the administration process of Clean Power Plan

i. *"we believe epa's guidelines exceed the authority congress intended in the clean air act, and do not follow three principles of our nation's laws"*

ii. *"Holding only one hearing deprives countless Americans of the opportunity to present their views in person to EPA -an opportunity EPA is required to provide under the Clean Air Act."*

iii. *"underscored by his ignoring the requests of fourteen states and cities for additional hearings so that our residents can voice their concerns to the agency directly."*

iv. *"I'm also requesting that you do additional hearings throughout the country and not just do one here in Charleston."*

v. *"Second, the CPP violated the concept of cooperative federalism that is a bedrock principle of the Clean Air Act and section 111(d) specifically"*

vi. *"Thousands of people testified in favor of the proposed Clean Power Plan in public hearings throughout the country"*

- Mention of ethical responsibility
  i. *"we have a duty to uphold a clean air environment not only to our citizens of the US but to the citizens of the World"*
- Mention of State's protected/violated rights, fairness/unfairness on applying CPP
  i. *"Stop stretching federal law to control and infringe upon states' rights."*
  ii. *"In Athens County, million gallons of toxic frack waste have been shipped in from Pennsylvania and West Virginia because Ohio has been granted primacy, where the other states have not."*

iii. *"take into account each state's capacities for reducing its use of coal, and it recognizes each state's discretion on the means for achieving those reductions"*

iv. *"each state can personalize and tailor the regulations so that it best meets the energy needs of that state"*

A.2.4.3 Stewardship

- Mention of creation, god's creation, stewardship
  i. *"care for creation"*
  ii. *"we are all called to be responsible stewards of the earth and to use the gifts we have been given to protect human life and dignity"*
- Mention of general concept of taking care of the environment/nature
  i. *"What's more, these mines produce the coal and then return the land as good, if not better than before mining began."*
  ii. *"truly embrace a century energy policy that respects all living things"*
  iii. *"we should treat the earth responsibly"*
- Not Stewardship: mere mentions of "protect the environment"

A.2.5 Multi-label topic coding example

Comment:

good evening. my name is rebecca blood, b as in becky, l-o-o-d. i'm here to speak on behalf of the national ***hydropower*** *association. national* ***hydropower*** *association, nha, is a nonprofit associate dedicated to representing the interests of the united states* ***hydropower***

industry. we include all technologies, conventional, pump storage, new marine, and hydrokinetic technologies, and our membership is vast. we're here to talk about america's leading ***renewable*** *electricity resource.* ***hydropower*** *provides approximately seven percent of our nation's total electricity supply, or , megawatts of installed capacity. and the majority of america's total* ***renewable*** *electricity.* ***hydropower*** *can double its contribution towards a* ***clean energy*** *economy, helping to avoid millions of additional tons of* ***carbon emissions****, serve as a* ***cost*** *effective compliance option, create hundreds of thousands of new* ***jobs****, strengthen our national infrastructure, and increase the reliability and resiliency of our electric grid, all priorities for the current administration. we appreciate this opportunity to talk to you today about the* ***clean energy*** *plan. we see this as an opportunity for federal and state regulatory decision makers to include* ***hydropower*** *as a viable compliance tool to reduce overall* ***emissions****. therefore, nha respectfully asks that epa and states consider the following recommendations. one, that all existing and future* ***hydropower*** *resources and technologies, should be recognized as compliance options.* ***hydropower*** *technologies include conventional, technologies including incremental* ***hydropower****, retrofits, capacity additions, and efficiency upgrades, pump storage, marine and hydrokinetics, conduits, and new development. our second recommendation is that epa's guidance to states should recognize existing programs and activities, and ensure the use of* ***hydropower*** *toward meeting* ***emission*** *reduction goals. and number three, epa should afford states maximum flexibility in designing their state implementation plans in this proceeding. we've reviewed the existing draft guidance, and we're evaluating additional impacts. we want to emphasize that what we've seen so far is that we are concerned whether the rule is actually recognizing and taking into account all the* ***clean***

air *benefits of* ***hydropower*** *and how it provides as well as providing the necessary market signals in support of new* ***hydro*** *development. so we are analyzing these issues in the rule, and we anticipate that we will file comments on the treatment of* ***hydropower*** *in this debate, and we want to determine a chance for states to look at us as a compliance option, and how it will be used by each state and by each region throughout this country. so thank you for this opportunity. we're very grateful for this time to bring these issues before your attention. mr. niebling*

Assigned topics: ***Climate, Future Economy, Pollution, Renewables, Cost, Jobs***

APPENDIX B. HUMAN ANNOTATOR TRAINING GUIDE: LABELING SENTIMENT AND TOPICS OF USER GENERATED REVIEWS ON ELECTRIC VEHICLE CHARGING EXPERIENCE FOR SUPERVISED MACHINE LEARNING

B.1 Research Objectives:

The objective of this research is to train a supervised machine learning model to identify the sentiment and the discussed topic from user generated text. The purpose of this training manual is to train annotators to follow established rules for consistent ratings to achieve high inter-rater reliability.

B.2 Labeling Tasks

B.2.1 Sentiment Labeling Task

- Identify whether the review presents negative or positive sentiment.
- The focus of the sentiment analysis is detecting negative sentiment, therefore if a review is not negative, then it should be labeled as positive.
- Negative Sentiment:
    - Notification of unavailability for successful charging
    - Expressing concerns
    - Any negativity overrides positivity

- Examples:
  - *"Out of order"*
  - *"A non ev car parking in the lot. can't get to fast charger."*
  - *"OUT OF SERVICE AGAIN! This station is a waste of time"*
  - *"Never lucky enough to get a spot to charge, someone's always there. Good luck!"*

- Positive Sentiment:
  - *Explicit positives*
  - Non-negative information sharing (confusing ones)
    - *"They have three charging stations right by the entrance."*
    - *"The other station is free now"*
    - *"Charged! When I called Blink CS before I traveled they said a tech had been here to fix this station, and I am happy to report it is!"*
    - *"Huge solar panels power this amazing station!!"*
    - *"Surprisingly not ICEd at 5:45pm on a Tuesday. Stall2B seemed slow, delivering only 28 KW at 45% SOC .moved to 3A."*

B.2.2 Main Topic Labeling Task

There are 9 topics to label, which are:

Topic	Subtopic
Functionality	Screen, Charger & Power Level, Connector Type, Card/Card Reader, Connection, Error Message, Mobile Application, Customer Service, Transaction, Safety, Other Functionality
Station Availability	# of Stations Available, ICE, General Congestion
Communication to other users	Charger Etiquette, Anticipated Time Available, User Tips
Location Features	General Location/Accessibility, Directions, Staff/people, Amenities, Points of Interest, Location Safety, Signage, Operation Hours
Cost	Parking, Charging
Range Anxiety	Trip, Range
Charging Speed	Charging Speed
Dealership	Dealership Charging Experience, Competing Brand Quality, Relationship with Dealers
Other	General Experiences

Reviews often fall into multiple topics. They are independent labels, and multiple topics can occur in the same review. Please select all the topics you think the review is discussing. However, Others topic is mutually exclusive with the rest of the topics. If any of the Functionality, Station Availability, Communication to Other Users, Location Features, Cost, Range Anxiety, Charging Speed, and Dealership topics are selected, Others should never be selected. When none of the 8 topics were selected, Others must be selected.

B.2.3 Multi-label Examples

"This QuickCharger does seem like it's a Level 2.5. :) It may terminate charging prematurely, in which case you'll have to contact Greenlots to give you a free follow-up session. Also, the highway exits to and from this charger location are confusing with the

one way streets and especially now with the construction. Go slow and follow signs, not necessarily your GPS."

- Functionality – mention of charger type (QuickCharger) and network (Greenlots)
- Service Time – mention of charging speed (*QuickCharger does seem like it's a Level 2.5, terminate charging prematurely)*
- Location – mention of location (*highway exits to and from this charger location are confusing*)
- User Interaction – givin advice to other users (*Go slow and follow signs, not necessarily your GPS*)

B.2.3.1 Functionality Topic

Functionality refers to comments describing whether or not particular features or services are working properly. Comments regarding station functionality are typically negative, as locations often face issues in any one of the given sub-topics above. If the charging capabilities of these charging stations are impaired in any way, users cannot achieve the goal of successfully charging their vehicles.

- Functionality:
    - Discussion on the following topics:
        - *Other Functionality, Charger, Screen, Power Level, Connector Type, Card, Reader, Connection, Time, Error Message, Station, Mobile, Application, Customer Service*

- Charger & Power level
    - *many of the chargers are down*
    - *Using a Supercharger. 3 waiting behind me.*
    - Mention of Level 1 (L1), Level 2 (L2), DC Charger, Fast Charger, Quick Charger (QC)
    - *qc slows down significantly after 20 min of course you are welcome to charge full 30 min if there is no car waiting*
- Screen:
    - *2 chargers have screens broken third could not connect called company and still no success*
- Connector type:
    - J1772, CHAdeMO, SAE, IEC Type 2, CCS, Tesla, Wall outlet, NEMA plug,
    - Tesla Model S, Tesla Supercharger, Tesla Roadster
    - *"Great friendly staff. Staff parks their cars in the two Tesla HPWC and one J1772 spaces to reserve these spots for EVs."*
- Card & Card reader
    - *"Chargepoint card required."*
    - *"Didn't realize needed ChargePoint Card"*
    - *3 of 4 units have issues either completely broken or card reader and lcd touchscreen does not work*
- Connection:
    - *dc charger could not secure connection*

        - *2 chargers have screens broken third could not connect called company and still no success*
        - *Second post now online*
    - Transaction:
        - *I mind the 10 minute over-the-phone transaction*
    - Error message
        - *"Error"*
        - *temperature error requires admin assistance can't charge*
        - *error on screen*
    - Mobile Application:
        - *i went onto the charge point app and got it started*
        - *needed a charge point account to unlock. easy to do with the app*
    - Customer Service
        - Getting help about functionality issues from customer service
        - *right hand charger not working called ipl to report left hand charger currently in use by another volt*
        - *customer service could not boot the station*
        - *2 chargers have screens broken third could not connect called company and still no success*
    - Safety: Mention of technical safety (not locational safety- see Location Features)
        - *Plug is broken and cord is exposed.. Kinda scary..*

- *Still up & running both terminals stay safe all rubber gloves for plugs*
- *If the red button is pressed, it remains locked even if the button is retracted, which requires undoing the panels and lifting a safety device.*
- *We asked about charging, after a very long time they told us we couldn't charge because they hadn't decided if they should ask money for it or not. And for our own safety, because they don't know if it would be safe.*
- *Unit switched off for safety reasons.*
- *Currently out of service due to damage. Manx Utilities are aware and have disabled the CP for safety.*

- Other Functionality: Can identify whether the charger is working for the customer (counter examples will show under "Not Functionality" section). No specification is mentioned on the Functionality issues.
  - *"charging now", "great charging"*
  - *"charged"*
  - *"broken"*
  - *"could not charge",*
  - *"I needed extra charging! Thanks"*
  - *"Charging my Zero S while here for monthly meeting of our electric car club. Www.eevc.info"*

- Not Functionality:
    - Cannot confirm whether the charger is working for the customer of the review
        - *"a charged car was parked there"*
        - *"a charged Nissan"*
        - *"Got here at opening time for Country Cafe. Very easy charge and location. Great spot for top-ups."*
        - *"Spot was ICE'd. Guess I'll be level 2 charging for the next 2 hours in 17 degree weather to get home."*
    - Limited operation due to accessibility
        - *"Not operable because the electricity is cutoff after 8pm"*
        - *"Could not charge because the dealership was closed"*
    - Does not specify the functionality issue
        - *"I hate coming here to charge. It is one of the worst places to charge. There is always an issue trying to charge here."* - Location

B.2.3.2 Station Availability Topic

Station Availability refers to comments concerning whether chargers are available at a given station. Users often comment about how many other cars are charging and how many chargers are not in use. This is an important aspect of the electric vehicle owners' experiences as they cannot successfully charge their vehicles without the availability of the chargers to do so. Users will also let the community know typical busy or slow times for

certain chargers with phrases like “busy on Saturday night” or “all chargers available Sunday morning”.

- Station Availability:
    - Number of stations available
        - *“2 out of the 4 chargers were down”*
        - *“there are total of XX chargers”*
        - *“There are many spots”*
    - Mention of on non-electric vehicle taking up EV charging parking spot (ICE)
        - *“Ice” mentions – “ICED”, “Ice’d”, “Icing”, “ice”*
        - *“Charged here, very convenient for shopping at a Target, wish all of them had a station. The other spot was taken by a jack A$$ who owns an ICE Ford piece of junk, does not know how to read!!!”*
        - *“non-ev was blocking the spot”*
    - General Congestion:
        - Information on occupation of charging spots
        - *“two spots are free now”*
        - *“There is long line for the charger”*
        - *“busy on Saturday night”*
        - *“all chargers available Sunday morning”*
        - *Using a Supercharger. 3 waiting behind me.*

- Not Availability:
    - No specific information of number of spots or chargers
        - *"the left one worked, right one didn't"*
        - *"one on far left and one on far right"*
    - Mention of specific location within charging space (See Location Features)
        - *"there is a spot at the right corner of the building"*
        - *"spot on the 2nd floor of the parking deck"*
    - Mention of the time when a space will be available (See Communication to other users)
        - *"you can unplug me at 12pm"*
        - *"I will be leaving at 6pm"*

B.2.3.3 Communication to other users

User interaction refers to comments in which users are directly interacting with other vehicle owners in the community. This is a unique topic in that users are conversing with or commenting to other users directly. Electric vehicle owners often refer to particular cars when asking to be plugged in or plugged out. This can also include questions being asked or user tips for a particular station or area.

- Help or Communication to other users:
    - Anticipated Time Available
        - *"Here for 1 hr 30 mins."*

        - *"Please plug me out after 90 minutes"*
        - *"1045 am charging for one hour with our jesla evse on way to quick charge in lancaster"*
        - *"all chargers on on both floors in use saturday aug 29 left notes for chargers on 5th floor to please plug me in when done thank you volt owner for doing that"*
        - *"sorry gray volt and thanks for leaving a note allowing me to unplug you i'll be sure to plug you in when i have enough charge to get home i was bone dry"*
    - User Tips: Seeking/Giving feedback or advise from/to other users
        - *"i might take a risky drive this weekend from south burlington to rutland to try out this station if anyone can confirm anything about it being up and running that might save me a headache"*
        - *"l3 still down now 8 of 9 fast chargers on 101 corridor are down creating an easy valley ev charging crisis mechanic is scheduled now so i am waiting around to ask questions about why so many are down so long"*
        - *"Can somebody post a photo of the new connector? I hope it's the new and improved V1.."*
        - Giving feedback or advise to others
        - *"stephen you can use the greenlots app if you don't have a fob"*

        - *"we will be here for a couple of hours we used our blink incard & forgot to check the responsiveness of the touch screen no issues with the card though 😊"*
        - *"charged successfully but you need to wiggle and push in handle manually the lever is broken av is notified"*
        - *"thanks for the tips on using the blink app which i now have i also called blink to report the broken touch screen"*
    - Charger Etiquette
        - *topped up my (attended) volt for the trip home it's so generous of mitsubishi to continue to open their chargers to the community however i noticed a silver volt owner leave their car and bike away not cool"*

B.2.3.4 Location Features Topic

Reviews that discuss general location information, directions, staff, description of amenities, point of interest, user activity while charging, signage, operating hours belong to this topic. The location topic refers to comments about various features or aspects specific to a particular charging station location. Users are interested in the amenities and features of the stations they visit before arriving there to charge. This community provides one another with helpful tips and tidbits about which locations have the best food, staff, or surrounding businesses. These reviews ensure users have the best knowledge about the

stations at which they are charging their vehicles and ensure that users are most prepared for their charging experiences.

- Location Features:
    - General Location/Accessability
        - *"Great spot"*
        - *"A very handy location! It's just In range of Fischer in Titusville. There are two chargers here BTW..."*
        - *"Perfectly located. Came 70 miles from Rochdale. GOM still showed 37 miles when I got here. Brattleboro next."*
    - Directions: Address, GPS discussions, instructions to be at the station
        - *"Charge station worked fine. Glad is at foot of mountain. My range anxiety went down a notch :)"*
        - *"address?"*
        - *"The GPS took me to a wrong place"*
        - *"There is a spot at the right corner of the building"*
        - *"spot on the 2nd floor of the parking deck"*
        - *"This QuickCharger does seem like it's a Level 2.5. :) It may terminate charging prematurely, in which case you'll have to contact Greenlots to give you a free follow-up session. Also, the highway exits to and from this charger location are confusing with the one way streets and especially now with the construction. Go slow and follow signs, not necessarily your GPS."*
    - Staff/people

- *"staff is friendly"*
- *"the people are so nice here need a urgent charge they are just great people here"*
- *the fast charger is not powered up, spent 9 hours with the dealership today L2 charging on both sides of a trip to N cal. staff is nice. got to talk a lot of the leaf. lots of miss information other in the sticks.*

- Amenities
  - *We rented an Airstream trailer overnight here and they let us charge our Telsa Model S overnight for free. Great amenities see their website. Easy walking distance to the Circus Circus Resort Casino.*
  - *Bathroom, hand wash,*
- Points of Interest
  - *We rented an Airstream trailer overnight here and they let us charge our Telsa Model S overnight for free. Great amenities see their website. Easy walking distance to the Circus Circus Resort Casino.*
  - *in hotel room 115 checked in at 9:30 pm hotel said good to leave car in charger*
- Location Safety
  - *great place and safe parking house. Did get 53km per hour.*
  - *"Free and safe location, better than commuting parking."*

        - *I have charged my Tesla X here at numerous occations. My car is always safe and the spot is always available. You need a red adapter!*
        - *Nice safe charger behind hotel steel gate*
        - *"I am glad! I connected the car for 20h, it was fully charged, no problems with the service, the car was safe all night "*
    - Signage
        - *Along with two others. Still 3 spots open! Two are labeled 45 min only.*
        - *"Two hour limit stated on sign."*
    - Operation Hours
        - *"Not operational after 8pm. The electricity powers off."*

B.2.3.5 Cost Topic

Reviews that discuss parking, charging fees and payment belong to Cost topic. Pricing is another big concern in the electric vehicle community. The pricing topic refers to comments about the amount of money required to park and/or charge at particular locations. The electric vehicle community is excited by free charging locations and readily shares praise surrounding the free locations.

- Cost: Overall cost that are required to charge

- Cost for charging
    - *"Charged for free"*
    - *"expensive"*
    - *"it is free"*
    - *"charging is complementary"*
    - *"valet park for parking charge"*
- Cost for parking
    - *"charging is free, but you need to pay for parking"*
    - *$12 to park and $2.40/ hour to charge at my max 3.3kWh draw, def not cheaper to drive an EV. Wide adoption is a long way off if this keeps up.*

- Not Cost:
    - When "free" used as "available" in the context
        - *"there is a free spot"* - Availability

B.2.3.6 Range Anxiety Topic

Range anxiety refers to comments regarding EV users fear of running out of fuel mid-trip. Range confidence, therefore, refers to comments concerning routes and tactics of EV users confident in their vehicles ability to reach destinations of interest. Because EV charging stations are usually less common to find than traditional gas stations, this concern is one of the biggest barriers for wider EV adoptions. This topic entails user experiences related to their level of confidence in having sufficient amount of power before arrival at a destination.

- Range Anxiety:
    - Mention of travel/trip
        - *“excellent stop on the way to atlanta charging at 28/hr”*
        - *" looked good hooked up and got the red alarm light evgo couldn't reset both chargers not working - again this station needs to be replaced it is a lemon and it is the first station coming up coast hw1”*
        - *"greenlots app reports "offline" need to use l3 sae for "ev" trip otherwise skipping columbus”*
        - *“rad on our way to leavenworth”*
        - *“return trip from manzanita via av dcqc at cannon beach 8 miles on gom @ lbw 33 & snowing over the pass charger worked great”*
    - Mention of battery life (range)
        - *“got 145mph of range with my tesla and the chdemo adapter they also have a j1772 level 2 charger and some 11v outlets (but no nema 14-5 outlets)”*
        - *" went to a redwood symphony event at the theater this evening parked in upper lot and was delighted to see 4 schneider / chargepoint spots only 2 of the 4 were operational the left hand sides were both faulted cables plenty long to reach either side charge reached 100% about 20 minutes before i drove home thanks canada college and chargepoint”*

        - *"20% to 80% in 25 min on my leaf however unlike the other quick chargers that raise my battery temperature another bar up this one did not that's great for my long trip"*
        - *"thank you curtis consulting i am down to 1 bar 7 miles on the gom 10 miles and a big hill to home and hungry you saved us"*
    - Mention of charge need (range)
        - *"needed a boost in charge and found this place thank you fork lift central"*
        - *"working great what a lifesaver i thought i' d be trapped out here far away from home"*
        - *"scanned card and screen blipped out stranded in seattle"*

B.2.3.7Charging Speed Topic

The service time topic refers to comments reporting charging rates experienced in a session. These comments typically consist of only the statistics given in unites of mileage or kilowatts per hours. Other units would include mentioning voltage or charging speed achieved (e.g. fast charge).

- Charging Speed:
    - Put list of common abbreviation (kw, v, amp,)
    - Mention of charging speed
        - *"slow charge" / "charged very fast"*
        - *"Reached 50miles in 10 minutes"*

        - *"Charging at 30mi/hr"*
    - Mention of voltage
        - *"198V at 30A"*
    - Mention of electrical power, current
        - *"90KW on far left"*
        - *"Peak charging power: 30amps"*
        - *"No ICE issues. One other model S at the 50A station, other two were open"*
- Not Charging Speed:
    - No specific mention of the charging speed
        - *"I charged here for 10 mins and left"* - Functionality
    - Mention of service hours of the charging station
        - *"plugs only on during business hours they are free although they salespeople will try to 'trade u outta' your car lol"* - Location
    - Mention of limited use time
        - *"along with two others still 3 spots open two are labeled 45 min only"* - Availability, Location

B.2.3.8 Dealership Topic

The dealerships topic refers to comments concerning specific dealerships and user's associated charging experiences. These comments serve an important source of information regarding a major stakeholder relationship influential to electric vehicle policy

making. Electric vehicle owners' feelings in this subtopic are largely determined by the different dealerships' accessibility regarding public charging.

- Dealership:
    - Dealership charging experience
        - *"on left side of building friendly Nissan dealer"*
    - Competing brand quality
        - *"These Nissan chargers really suck balls. Temp error and the guard is in control of the breaker. making me wait 20 minutes before he'll turn the power back on. Really wishing I had a Tesla right about now"*
    - Relationship with the dealers
        - *"thank you very nice dealership"*
    - Mention of dealers
        - *"Car dealers please note: new drivers should get a lesson on how to use these chargers as they are not intuitive and new drivers have broken the connectors previously at some of the stations because nobody has shown them how to use these connectors."*
        - *"very easy to find charging stations are by the front doors of the dealership"*
        - *"chademo is still free but requires a chargepoint card if you don't have one the dealer will use theirs for your charging session"*
- Not Dealership
    - No specific mention of the car name or brands, or the word "dealer"

        - *"awesome little store and chris is super nice and friendly"* - more like Location (staff)
    - Mere mention of car names and brands
        - Leaf (Nissan), I3 (BMW), Tesla Model S, Volt (Chevolet), 500e (Fiat), Spark (Chevolet), C-max Energi (Ford), Fusion Energi (Ford), Prius (Toyota), RAV4 EV (Toyota), Soul (Kia)
        - *"quick charge working great salesman came out and turned it on for us as we aren't in the network thanks magic Nissan"*
        - *"first time i have seen a full house at the palm street parking garage three chevy volts and one nissan leaf"*
        - *"saw the silver volt in the right side spot i think he/she just wants that spot not really needed to charge sad"*
        - *"great to see a tesla s visiting the car show at the hotel"*

REFERENCES

Alvarez, K., Dror, A., Wenzel, E., and Asensio, O. I. (2019). Evaluating electric vehicle user mobility data using neural network based language models. In Proceedings of the 98th annual meeting of the Transportation Research Board.

Anderson, J. E., Lehne, M., and Hardinghaus, M. (2018). What electric vehicle users want: Real-world preferences for public charging infrastructure. International Journal of Sustainable Transportation, 12(5):341–352.

Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., & Ha, S. (2020). Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. Nature Sustainability, 3(6), 463–471. https://doi.org/10.1038/s41893-020-0533-6

Asensio, O. I., Alvarez, K., Dror, A., Wenzel, E., Hollauer, C., and Ha, S. (2020) Real-time data from mobile platforms to evaluate sustainable transportation infrastructure. Nature Sustainability, 3:463–471.

Asensio, O. I., Mi, X., and Dharur, S. (2020). Using machine learning techniques to aid environmental policy analysis: a teaching case in big data and electric vehicle infrastructure. Case Studies in the Environment, 961302.

Axsen, J., Langman, B., and Goldberg, S. (2017). Confusion of innovations: mainstream consumer perceptions and misperceptions of electric-drive vehicles and charging programs in canada. Energy Research & Social Science, 27:163–173.

Bell, S. E., & York, R. (2010). Community Economic Identity: The Coal Industry and Ideology Construction in West Virginia. Rural Sociology, 75(1), 111–143. https://doi.org/10.1111/j.1549-0831.2009.00004.x

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Brückmann, G. M., and Bernauer, T. (2020). What drives public support for policies to enhance electric vehicle adoption? Environmental Research Letters.

Burgess, M., King, N., Harris, M., and Lewis, E. (2013). Electric vehicle drivers' reported interactions with the public: Driving stereotype change? Transportation Research Part F: Traffic Psychology and Behavior, 17:33–44.

California Renewables Portfolio Standard Program: Emissions of greenhouse gases., Pub. L. No. SB100 (2018). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201720180SB100

Carley, S., Krause, R. M., Lane, B. W., and Graham, J. D. (2013). Intent to purchase a plug-in electric vehicle: A survey of early impressions in large us cites. Transportation Research Part D: Transport and Environment, 18:39–45.

Chargemap. Chargemap's community. (2020). Access date: 10/27/2020, https: //chargemap.com/community.

ChargePoint. Chargepoint map. (2020). Access date: 10/27/2020, https://na.chargepoint.com/charge_point.

Costa, M., Desmarais, B. A., & Hird, J. A. (2019). Public Comments' Influence on Science Use in U.S. Rulemaking: The Case of EPA's National Emission Standards. The American Review of Public Administration, 49(1), 36–50. https://doi.org/10.1177/0275074018795287

Cowgill, B., and Tucker, C. (2017). Algorithmic bias: A counterfactual perspective. In Workshop on Trustworthy Algorithmic Decision-Making.

Czajka, J. L., & Beyler, A. (2016). Declining Response Rates in Federal Surveys: Trends and Implications (40146.D4C; p. 86). Mathematica Policy Research.

De'Arman, K. J. (2020). Is Public Participation Public Inclusion? The Role of Comments in US Forest Service Decision-Making. Environmental Management, 66(1), 91–104. https://doi.org/10.1007/s00267-020-01278-5

Department of Energy. Electric vehicles: Tax credits and other incentives database. (2019). Access date: 07/31/2019, https://www.energy.gov/eere/electricvehicles/electric-vehicles-tax-credits-and-other-incentives.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). Bert: Pretraining of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).

Egbue, O., and Long, S. (2012). Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions. Energy Policy, 48:717–729.

Environmental Protection Agency. (2018). Greenhouse gas emissions from a typical passenger vehicle. document No. 420-F-18-008.

Environmental Protection Agency. (2018). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2016. document No. 430-R-18-003.

European Commission, Directorate-General for Mobility and Transport. (2011). White Paper on Transport: Roadmap to a Single European Transport Area: Towards a Competitive and Resource-Efficient Transport System. Office of the European Union.

European Parliament and Council of the European Union. (2014). Directive 2014/94/eu of the European parliament and of the Council of 22 October 2014 on the deployment of alternative fuels infrastructure text with EEA relevance. Official Journal of the European Union, 57(L307):1–20.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Godby, R., & Coupal, R. (2016). The potential impact of rate-based or mass-based rules on coal-producing states under the Clean Power Plan. The Electricity Journal, 29(6), 42–51. https://doi.org/10.1016/j.tej.2016.07.006

Godby, R., Coupal, R., Taylor, D., & Considine, T. (2015). Potential Impacts on Wyoming Coal Production of EPA's Greenhouse Gas Proposals. The Electricity Journal, 28(5). https://doi.org/10.1016/j.tej.2015.05.004

Grubert, E. (2017). How to Do Mail Surveys in the Digital Age: A Practical Guide. Survey Practice, 10(1), 2787. https://doi.org/10.29115/SP-2017-0002

Grubert, E. (2019). Every Door Direct Mail in US survey research: An anonymous census approach to mail survey sampling. Methodological Innovations, 12(2), 2059799119862104. https://doi.org/10.1177/2059799119862104

Grubert, E., & Siders, A. (2016). Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses. Environmental Research Letters, 11(9), 093001. https://doi.org/10.1088/1748-9326/11/9/093001

Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. PLoS ONE, 15(5). https://doi.org/10.1371/journal.pone.0232076

Ha, S., and Marchetto, D. J. (2020). Labeling sentiment and topics of user generated reviews on electric vehicle charging experience for supervised machine learning. https://github.com/asensio-lab/transformer-EV-topic-classification/blob/master/training-manual/training-manual.pdf.

Ha, S., Marchetto, D. J., Burke, M. E., and Asensio, O. I. (2020). Detecting behavioral failures in emerging electric vehicle infrastructure using supervised text classification algorithms. In Proceedings of the 99th annual meeting of the Transportation Research Board.

Ha, S., Marchetto, D. J., Dharur, S., & Asensio, O. I. (2021). Topic classification of electric vehicle consumer experiences with transformer-based deep learning. Patterns, 2(2), 100195. https://doi.org/10.1016/j.patter.2020.100195

Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., Figenbaum, E., Jakobsson, N., Jochem, P., Kinnear, N. et al (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure. Transportation Research Part D: Transport and Environment, 62:508– 523.

Haynes-Maslow, L., Andress, L., Jilcott Pitts, S., Osborne, I., Baquero, B., Bailey-Davis, L., Byker-Shanks, C., Houghtaling, B., Kolodinsky, J., Lo, B. K., Morgan, E. H., Piltch, E., Prewitt, E., Seguin, R. A., & Ammerman, A. S. (2018). Arguments Used in Public Comments to Support or Oppose the US Department of Agriculture's Minimum Stocking Requirements: A Content Analysis. Journal of the Academy of Nutrition and Dietetics, 118(9), 1664–1672. https://doi.org/10.1016/j.jand.2017.12.005

Hazboun, S. O., Briscoe, M., Givens, J., & Krannich, R. (2019). Keep quiet on climate: Assessing public response to seven renewable energy frames in the Western United

States. Energy Research & Social Science, 57, 101243. https://doi.org/10.1016/j.erss.2019.101243

Hemmatian, B., Sloman, S. J., Cohen Priva, U., & Sloman, S. A. (2019). Think of the consequences: A decade of discourse about same-sex marriage. Behavior Research Methods, 51(4), 1565–1585. https://doi.org/10.3758/s13428-019-01215-3

Hidrue, M. K., Parsons, G.R., Kempton, W., and Gardner, M. P. (2011). Willingness to pay for electric vehicles and their attributes. Resource and Energy Economics, 33(3):686–705.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8):1735–1780.

Hoekstra, A. (2019). The underestimated potential of battery electric vehicles to reduce emissions. Joule, 3(6):1412 – 1414.

Jung, M. F., Sirkin, D., Gür, T. M., and Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. ArXiv:1404.2188 [Cs]. http://arxiv.org/abs/1404.2188

Kam, M. V. D., Sark, W. V., and Alkemade, F. (2020). Multiple roads ahead: How charging behavior can guide charging infrastructure roll-out policy. Transportation Research Part D: Transport and Environment, 85:102452.

Kempton, W., Tomic, J., Letendre, S., Brooks, A., and Lipman, T. (2001). Vehicle-to-grid power: battery, hybrid, and fuel cell vehicles as resources for distributed electric power in California.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Krause, R. M., Carley, S. R., Lane, B. W., and Graham, J. D. (2013). Perception and reality: Public knowledge of plug-in electric vehicles in 21 us cities. Energy Policy, 63:433–440.

Kühl, N., Goutier, M., Ensslen, A., and Jochem, P. (2019). Literature vs. Twitter: Empirical insights on customer needs in e-mobility. Journal of Cleaner Production, 213:508–520.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In International Conference on Learning Representations.

Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1):159–174.

LeCun, Y., and Bengio, Y. (1998). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, pp. 255-258.

Lee, Z. J., Pang, J. Z. F., and Low, S. H. (2020). Pricing EV charging service with demand charge. Electric Power Systems Research, 189:106694.

Liao, F., Molin, E., and Wee, B. V. (2017). Consumer preferences for electric vehicles: a literature review. Transport Reviews, 37(3):252–275.

Liu, W., Zhou, P., Wang, Z., Zhao, Z., Deng, H., and Ju, Q. (2020). FastBERT: a self-distilling BERT with adaptive inference time. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Lynes, J. (2018). Dealerships are a tipping point. Nature Energy, 3(6):457–458.

Ma, B., Zhang, N., Liu, G., Li, L., & Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. Information Processing & Management, 52(3), 430–445. https://doi.org/10.1016/j.ipm.2015.10.004

Matthews, L., Lynes, J., Riemer, M., Matto, T. D., and Cloet, N. (2017). Do we have a car for you? Encouraging the uptake of electric vehicles at point of sale. Energy Policy, 100:79–88.

McCollum, D. L., Wilson, C., Bevione, M., Carrara, S., Edelenbosch, O. Y., Emmerling, J., Guivarch, C., Karkatsoulis, P., Keppo, I., Krey, V. et al. (2018). Interaction of consumer preferences and climate policies in the global transition to low-carbon vehicles. Nature Energy, 3(8):664– 673.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs]. http://arxiv.org/abs/1301.3781

Morstyn, T., Farrell, N., Darby, S. J., and McCulloch, M. D. (2018). Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. Nature Energy, 3(2):94–101.

National Research Council. (2010). Hidden costs of energy: unpriced consequences of energy production and use. National Academies Press.

National Research Council. (2015). Overcoming barriers to deployment of plug-in electric vehicles. National Academies Press.

Nguyen, A. T., Wallace, B. C., Li, J. J., Nenkova, A., and Lease, M. (2017). Aggregating and predicting sequence labels from crowd annotations. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

Nicolson, M., Huebner, G. M., Shipworth, D., and Elam, S. (2017). Tailored emails prompt electric vehicle owners to engage with tariff switching information. Nature Energy, 2(6):1–6.

Noel, L., and Sovacool, B. K. (2016). Why did better place fail?: range anxiety, interpretive flexibility, and electric vehicle promotion in denmark and israel. Energy Policy, 94:377–386.

Open Charge Alliance. Open charge point protocol 2.0.1 specification. (2020). Released: 03/31/2020.

Open Charge Alliance. Open smart charge protocol 2.0 specification. (2020).

Open Charge Map. Open charge map community. (2020). Access date: 10/27/2020, https://community.openchargemap.org/.

Rambachan, A., Kleinberg, J., Ludwig, J., and Mullainathan, S. (2020). An economic perspective on algorithmic fairness. In AEA Papers and Proceedings, 110:91–95.

Rauh, N., Franke, T., and Krems, J. F. (2015). Understanding the impact of electric vehicle driving experience on range anxiety. Human Factors, 57(1):177– 187.

Recargo. Plugshare key features and benefits. (2020). Access date: 02/13/2020, https://recargo.com/plugshare.html.

Recharge. United states EV charging network interoperability is a lie. (2020). Access date: 08/31/2020, https://www.evpassport.com/post/us-ev-charging-networkinteroperability-is-a-lie.

Roberson, L. A., and Helveston, J. P. (2020). Electric vehicle adoption: can short experiences lead to big change? Environmental Research Letters, 15(9):0940c3.

Rubens, G. Z., Noel, L., and Sovacool, B. K. (2018). Dismissive and deceptive car dealerships create barriers to electric vehicle adoption at the point of sale. Nature Energy, 3(6):501–507.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2):197–227.

Scott, T. A., Ulibarri, N., & Figueroa, O. P. (2020). NEPA and National Trends in Federal Infrastructure Siting in the United States. Review of Policy Research, 37(5), 605–633. https://doi.org/10.1111/ropr.12399

Serrano, S., and Smith, N. A. (2019). Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Shapiro, S. (2008). Does the amount of participation matter? Public comments, agency responses and the time to finalize a regulation. Policy Sciences, 41(1), 33–49. https://doi.org/10.1007/s11077-007-9051-x

Sheldon, L. T., DeShazo, J. R., and Carson, R. T. (2017). Electric and plug-in hybrid vehicle demand: lessons for an emerging market. Economic Inquiry, 55(2):695–713.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(56), 1929–1958.

Stern, M. J., Bilgen, I., & Dillman, D. A. (2014). The State of Survey Methodology: Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design. Field Methods, 26(3), 284–301. https://doi.org/10.1177/1525822X13519561

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Surowiecki, J. (2005), The wisdom of crowds. Anchor.

TEN-T. Eu-funded fast-charge network opens up pan-european travel for EV drivers. (2015).

US EPA, O. (2020). Electric Utility Generating Units: Repealing the Clean Power Plan [Other Policies and Guidance]. US EPA. https://www.epa.gov/stationary-sources-air-pollution/electric-utility-generating-units-repealing-clean-power-plan

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems.

Walsh, K. B., Haggerty, J. H., Jacquet, J. B., Theodori, G. L., & Kroepsch, A. (2020). Uneven impacts and uncoordinated studies: A systematic review of research on unconventional oil and gas development in the United States. Energy Research & Social Science, 66, 101465. https://doi.org/10.1016/j.erss.2020.101465

Wang, S., Wang, J., Li, J., Wang, J., and Liang, L. (2018). Policy implications for promoting the adoption of electric vehicles: Do consumer's knowledge, perceived risk and financial incentive policy matter? Transportation Research Part A: Policy and Practice, 117:58–69.

Wang, Y., Li, H., & Wu, Z. (2019). Attitude of the Chinese public toward off-site construction: A text mining study. Journal of Cleaner Production, 238, 117926. https://doi.org/10.1016/j.jclepro.2019.117926

Xu, Z., & Bengston, D. N. (1997). Trends in national forest values among forestry professionals, environmentalists, and the news media, 1982–1993. Society & Natural Resources, 10(1), 43–59. https://doi.org/10.1080/08941929709381008

Yan, F., Ruwase, O., He, Y., and Chilimbi, T. (2015). Performance modeling and scalability optimization of distributed deep learning systems. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Yang, H., & Callan, J. (2009). OntoCop: Constructing Ontologies for Public Comments. 24, 6.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems.

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923.

Zaidan, O. F., Eisner, J., and Piatko, C. (2008). Machine learning with annotator rationales to reduce annotation cost. In Proceedings of the NIPS* 2008 workshop on Cost Sensitive Learning.

Zhang, Y., and Wallace, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).